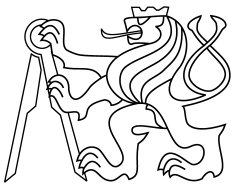
CENTER FOR
MACHINE PERCEPTION



CZECH TECHNICAL
UNIVERSITY IN PRAGUE

RESEARCH REPORT

ISSN 1213-2365

Revisiting the Decomposition Approach to Inference in Exponential Families and Graphical Models

Tomáš Werner

werner@cmp.felk.cvut.cz

CTU–CMP–2009–06

May 2009

This work was supported by the European Commission grant 215078 and the Czech government grant MSM6840770038.

Research Reports of CMP, Czech Technical University in Prague, No. 6, 2009

Published by

Center for Machine Perception, Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University
Technická 2, 166 27 Prague 6, Czech Republic
fax +420 2 2435 7385, phone +420 2 2435 7637, www: http://cmp.felk.cvut.cz

Revisiting the Decomposition Approach to Inference in Exponential Families and Graphical Models

Tomáš Werner

May 2009

Contents

1. Introduction

In a series of papers [18, 22, 20, 23, 21, 24], Wainwright and colleagues developed an approach to upper-bound the maximal probability (= the mode) and the partition function (and in turn, to approximate the marginals) of undirected graphical models (Markov random fields, MRF). It is based on decomposing the original problem into smaller subproblems with tractable structures. The formalism used actually allows for a wider class of probability distributions than the ones defined by a MRF – namely the general (discrete) exponential family of distributions. Despite this, the approach was detailed only for distributions defined by a MRF and only for tree-structured subproblems (though hypertrees were discussed too).

Our aim in this text is to generalize this approach to the general exponential family of distributions (rather than only those defining a MRF) and to general subproblems (rather than only (hyper)trees). In particular, we formulate and discuss in detail the upper bound minimization, its Lagrange dual, and the optimality conditions; outer bound of the mean polytope resulting from the decomposition is constructed; and a coordinate descent algorithm (similar to the Generalized Iterative Scaling) to optimize the upper bound. We do this both for non-zero temperature (i.e., the partition function) and the zero temperature limit (i.e., the modes), paying special attention to what is happening during the transition to zero temperature. The obtained results are simple and natural and reveal the full picture of the decomposition approach to statistical inference.

1.1. Exponential family of probability distributions

Consider the probability distribution

$$p(x\,|\,\boldsymbol{\theta}) = \exp[\boldsymbol{\theta}\boldsymbol{\phi}(x) - F(\boldsymbol{\theta})] \tag{1}$$

over a finite set X, i.e., $x \in X$. In (1), I is a finite set, $\boldsymbol{\phi}$ is a mapping $X \to \mathbb{R}^I$ with components ϕ_i, and $\boldsymbol{\theta} \in \mathbb{R}^I$ is a vector with components θ_i (for $i \in I$). Here, $\boldsymbol{\theta}$ is a row vector and $\boldsymbol{\phi}(x)$ a column vector[1], so that $\boldsymbol{\theta}\boldsymbol{\phi}(x) = \sum_{i\in I} \theta_i\phi_i(x)$. The normalization term

$$F(\boldsymbol{\theta}) = \bigoplus_{x\in X} \boldsymbol{\theta}\boldsymbol{\phi}(x) \tag{2}$$

is the *log-partition function*, where we introduced the symbol

$$\log(\mathrm{e}^a + \mathrm{e}^b) = a \oplus b \tag{3}$$

[1] Distinguishing row and column vectors emphasizes the fact that while vectors $\boldsymbol{\phi}(x)$ belong to the vector space $\mathbb{R}^I$, vectors $\boldsymbol{\theta}$ belong to the dual vector space and thus represent linear forms.

for the *log-sum-exp operation*. A distribution in form (1) is known as the (discrete) exponential family. The family is defined by triplet $(X, I, \boldsymbol{\phi})$ and parameterized by $\boldsymbol{\theta}$. The basis functions ϕ_i are usually referred to as *potential functions* or *sufficient statistics* and numbers θ_i as *canonical parameters*. It is also known as the *log-linear model* because the expression $\boldsymbol{\theta}\boldsymbol{\phi}(x)$ captures all functions that are linear in $\boldsymbol{\theta}$ for every x.

It is often of interest to calculate the log-partition function $F(\boldsymbol{\theta})$, the maxima (modes) of $p(x\,|\,\boldsymbol{\theta})$, and the mean values of the functions ϕ_i over $p(x\,|\,\boldsymbol{\theta})$. These are examples of inference tasks in the exponential family.

Clearly, x maximizes $p(x\,|\,\boldsymbol{\theta})$ if and only if it maximizes $\boldsymbol{\theta}\boldsymbol{\phi}(x)$, thus to find the mode it suffices to evaluate the function

$$\bar{F}(\boldsymbol{\theta}) = \max_{x \in X} \boldsymbol{\theta}\boldsymbol{\phi}(x) \tag{4}$$

The value of $\bar{F}(\boldsymbol{\theta})$ can be alternatively expressed as a limit of the log-partition function because

$$\bar{F}(\boldsymbol{\theta}) = \lim_{t \to 0+} tF(\boldsymbol{\theta}/t) \tag{5}$$

Equality (5) follows from the property of the log-sum-exp function that $\lim_{t\to 0+} t \bigoplus_k (a_k/t) = \max_k a_k$.

In statistical mechanics, the limit $t \to 0+$ is known as the *zero-temperature limit*. Distribution of type (1) was proposed there as a statistical description of a system composed of a large number of locally interacting parts. If $-\boldsymbol{\theta}\boldsymbol{\phi}(x)$ is the energy of the system as a function of its state x, the celebrated result by Boltzmann says that the probability that the system with temperature t is in state x equals $p(x\,|\,\boldsymbol{\theta}/t)$. When the temperature approaches zero, only the states that maximize $p(x\,|\,\boldsymbol{\theta})$ have a non-zero probability $p(x\,|\,\boldsymbol{\theta}/t)$. These so called *ground states* describe the 'frozen' system (such as crystals).

A possible instantiation of the exponential family (1) is a discrete *undirected graphical model* (*Markov random field*). Here, $p(x\,|\,\boldsymbol{\theta})$ is a joint distribution of a multiple discrete variables $x = (x_1, x_2, \ldots)$ and the functions ϕ_i are 0-1 functions such that their mean values coincide with the marginals of $p(x\,|\,\boldsymbol{\theta})$ on a chosen set of subsets of the variables.

1.2. Upper bounds from decomposition

If the set X is combinatorially large (such as for MRFs), evaluating functions F and $\bar{F}$ may be intractable. Then we are naturally interested in approximations, upper bounds, and tractable subclasses. One of existing approaches to obtain these, due to Wainwright et al. [18, 22, 20, 23, 21, 24], is based on *decomposition to subproblems* and its essence is as follows. Let $\{\, \boldsymbol{\theta}^s \mid s \in S \,\}$ be a collection of parameter vectors and let $\rho^s \geq 0$ be scalars such that $\sum_{s\in S} \rho^s = 1$. Since F is convex, applying Jensen's inequality to it yields

$$F\Big(\sum_{s\in S} \boldsymbol{\theta}^s\Big) \leq \sum_{s\in S} \rho^s F(\boldsymbol{\theta}^s/\rho^s) \tag{6}$$

i.e., the right-hand side is the upper bound in the left-hand side. If the subproblems $\boldsymbol{\theta}^s$ are such that evaluating $F(\boldsymbol{\theta}^s)$ is tractable for each s (typically, due to restricting the structure of the subproblems), this bound is tractable to compute. The upper bound is minimized over the collections $\{\boldsymbol{\theta}^s\}$ subject to the constraint that $F(\boldsymbol{\theta}^s)$ are tractable to compute and that $\sum_{s\in S} \boldsymbol{\theta}^s$ defines our original distribution. This leads to a smooth convex optimization task, provided that ρ^s are fixed. Including ρ^s in the optimization leads to a non-convex task and we do not consider it in this text.

For zero temperature, we have even a simpler inequality,

$$\bar{F}\Big(\sum_{s\in S} \boldsymbol{\theta}^s\Big) \leq \sum_{s\in S} \bar{F}(\boldsymbol{\theta}^s) \tag{7}$$

where ρ^s canceled out because $\bar{F}(\boldsymbol{\theta}/\rho) = \bar{F}(\boldsymbol{\theta})/\rho$. Minimizing the upper bound (7) leads to a non-smooth convex minimization task, in fact, to a linear programming.

After its minimization, the upper bound (6) is tight (i.e., holds with equality) only in trivial cases. In contrast, the least upper bound (7) is tight for a large and highly non-trivial class of instances, forming thus a tractable subclass of problem (4). Precisely, (7) holds with equality if and only if there exists $x^* \in X$ such that

$$x^* \in \operatorname*{argmax}_{x\in X} \boldsymbol{\theta}^s\boldsymbol{\phi}(x) \qquad \forall s \in S \tag{8}$$

i.e., if the subproblems agree on a common solution, x^*. This is a generalization of (hyper)tree agreement [18, 20].

1.3. Contributions

In this text, we are primarily interested in upper bounding the function $\bar{F}$ (i.e., the modes of $p(\mathbf{x}\,|\,\boldsymbol{\theta})$). Upper bounding the log-partition function F is of only secondary interest for us – we are interested in it mainly to shed light on the transition to zero temperature. We nevertheless derive it in full generality.

We formulate the decomposition approach for the general form of the exponential family (1). We do this first for the non-zero temperature case and then for the zero temperature limit. This allows us to see similarities and differences between them. In particular, we formulate

- **the minimization of the upper bound**

  We formulate the problem of finding the least upper bound on the mode and log-partition function for general subproblems and the general exponential family.

- **its Lagrange dual**

  We derive the dual of these tasks. This dual is very similar to the variational expression for the log-partition function and the mode, in which we maximize a concave function over the mean/marginal polytope: the mean polytope is replaced by its outer bound (obtained as the intersection of projections of the true mean polytope onto subproblems) and the true entropy term is replaced with the convex combination of the entropies of subproblems.

- **optimality conditions**

  For non-zero temperature, the primal-dual pair is jointly optimal iff the mean parameters of the overlapping subproblems coincide. For zero-temperature, optimality is characterized by a non-empty intersection of the optimal faces of the mean polytopes of the subproblems.

- **a coordinate descent algorithm to decrease the upper bound**

We present a coordinate descent algorithm to decrease the upper bound, motivated by the algorithms based on averaging (max-)marginals in MRFs (such as max-sum diffusion or TRW-S) and by the Generalized Iterative Scaling. We analyze its stationary points for the zero temperature limit.

- **a cutting plane algorithm to improve the bound incrementally**

  We sketch a cutting plane algorithm, which allows to improve the upper bound incrementally by adding more and more complex subproblems.

As an example, we apply the theory to pairwise MRFs, giving examples of typical collection of subproblems (trees, individual nodes and edges, cycles). Despite emphasizing that we develop the theory for the general form of the exponential family, we did not apply it to any practical problem that is not a graphical model – we could not find any such application. Although this deems our contribution to be of only theoretical importance, we believe such applications exist.

2. Notation

Sets are denoted by $\{\cdot\}$, ordered tuples by $(\cdot)$, intervals by $[a, b]$. The set of reals, non-negative reals and positive reals is $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{R}_{++}$, respectively. The set of all k-element subsets of set A is denoted by $\binom{A}{k}$. For a set A, we denote $f(A) = \{\, f(x) \mid x \in A \,\}$. Symbol $[\![\alpha]\!]$ equals 1 if expression α is true and 0 if α is false. Symbol $\operatorname{argmax}_x f(x)$ denotes the set of *all* maximizers of f.

For $A \subseteq \mathbb{R}^n$, $\operatorname{aff} A$ denotes the affine hull of A, $\operatorname{conv} A$ the convex hull, $\operatorname{ri} A$ the relative interior, and $\operatorname{rbd} A = A \setminus \operatorname{ri} A$ the relative boundary [3].

3. Properties of exponential family

This section surveys the properties of discrete exponential families we will need. In that, we mostly follow [21] and we occasionally use also [4, 1]. We give no formal proofs but most of the statements are easy to verify by elementary algebra. We pay special attention to overcomplete representation and reparameterizations (§3.2), which are crucial in context of MRFs – our treatment of these is more principled than in [21, 19]. We further include a note on Generalized Iterative Scaling (§3.5).

Let

$$\mathbf{Y} = \boldsymbol{\phi}(X) = \{\, \boldsymbol{\phi}(x) \mid x \in X \,\} \tag{9}$$

denote the map of X under $\boldsymbol{\phi}$, a finite set of vectors $\mathbf{y} \in \mathbb{R}^I$. Further in §3 and §4, we will adopt the following simplification: instead of the distribution $p(x\,|\,\boldsymbol{\theta})$ over set X given by (1) we will consider the distribution

$$q(\mathbf{y}\,|\,\boldsymbol{\theta}) = \exp[\boldsymbol{\theta}\mathbf{y} - F(\boldsymbol{\theta})] \tag{10}$$

over $\mathbf{Y}$, obtained by substituting $\mathbf{y} = \boldsymbol{\phi}(x)$. This simplifies the exposition but does not otherwise change the situation because X and $\boldsymbol{\phi}$ serve only to index the elements of $\mathbf{Y}$. If the mapping $\boldsymbol{\phi}$ is one-to-one, this substitution does not change the log-partition function,

$$\bigoplus_{x \in X} \boldsymbol{\theta}\boldsymbol{\phi}(x) = \bigoplus_{\mathbf{y} \in \mathbf{Y}} \boldsymbol{\theta}\mathbf{y} \tag{11}$$

We assume it is so. If not, the theory in in §3 and §4 could be easily restated also without this substitution. While the family of distributions $p(x\,|\,\boldsymbol{\theta})$ is defined by the triplet $(X, I, \boldsymbol{\phi})$, the family of distributions $q(\mathbf{y}\,|\,\boldsymbol{\theta})$ is defined by the pair $(\mathbf{Y}, I)$ where I is a finite set and $\mathbf{Y}$ is a finite set of vectors from $\mathbb{R}^I$.

3.1. Log-sum-exp operation

Exponential families have a number of deep and interesting properties. Many of them follow already from properties of the *log-sum-exp operation* [3]. To emphasize this fact, in (3) we introduced a special symbol for it, $\oplus$. This section summarizes its key properties.

The operation $\oplus$ is associative and commutative, thus it makes sense to write $\bigoplus_k a_k$. It is also distributive w.r.t. addition, $\bigoplus_k (b + a_k) = b + \bigoplus_k a_k$. Thus, $(\mathbb{R}, \oplus, +)$ is a commutative semiring (in fact, a semifield). It is isomorphic with the commutative semiring $(\mathbb{R}_{++}, +, \times)$ via the map $a \mapsto \log a$.

The mapping $\mathbf{a} \mapsto \bigoplus_k a_k$ is convex. In fact, it is the convex conjugate [3] of entropy.

Its derivative is

$$\frac{\partial}{\partial a_j} \bigoplus_k a_k = \exp\Big(a_j - \bigoplus_k a_k\Big) = \frac{\exp a_j}{\sum_k \exp a_k} \tag{12}$$

Since the numbers (12) sum up to 1, they can be seen as a probability distribution, which is often called the *soft maximum*.

Consider the family of functions $t \bigoplus_k (a_k/t)$, parameterized by $t \neq 0$. For any $t < 1$, we have

$$\bigoplus_k a_k > t \bigoplus_k (a_k/t) > \lim_{t \to 0+} t \bigoplus_k (a_k/t) = \max_k a_k \tag{13}$$

The limit in (13) is known as Maslov's dequantization or tropicalization [14, 8]. It takes the semiring $(\mathbb{R}, \oplus, +)$ into the max-sum (= tropical) semiring $(\mathbb{R}, \max, +)$. Informally, we can imagine this as mechanically replacing all occurencies of operations $\oplus$ in an expression with $\max$. In the logarithmic domain, it takes the sum-product semiring $(\mathbb{R}_{++}, +, \times)$ into the max-product semiring $(\mathbb{R}_{++}, \max, \times)$. Here, (13) states the well-known fact that the ℓ_p vector norm becomes the max-norm as $p \to \infty$.

The similar limit for (12) reads

$$\lim_{t \to 0+} \frac{\exp(a_j/t)}{\sum_k \exp(a_k/t)} = \begin{cases} 0 & \text{if } j \notin K^* \\ |K^*|^{-1} & \text{if } j \in K^* \end{cases} \tag{14}$$

where $K^* = \operatorname{argmax}_k a_k$. This equality is easily verified by computing the limit.

3.2. Overcomplete representation

If there are no affine dependencies among the vectors from $\mathbf{Y}$ (in other words, $\operatorname{aff} \mathbf{Y} = \mathbb{R}^I$) then $(\mathbf{Y}, I)$ is a *minimal representation* of the family (10). If the elements of $\mathbf{Y}$ are affinely dependent, it is an *overcomplete representation*. Then there is at least one vector $\boldsymbol{\theta}$ such that $\boldsymbol{\theta}\mathbf{y}$ is constant for all $\mathbf{y} \in \mathbf{Y}$. If this constant is zero, we speak about a homogeneous dependency. If this constant is non-zero, we speak about an inhomogeneous dependency – in that case, $\boldsymbol{\theta}$ can be scaled to make this constant equal to 1. Stacking all such vectors $\boldsymbol{\theta}$ as matrix

rows yields

$$\mathbf{A}\mathbf{y} = \mathbf{0}, \quad \mathbf{B}\mathbf{y} = \mathbf{1} \qquad \forall \mathbf{y} \in \mathbf{Y} \tag{15}$$

Matrix $\mathbf{A} \in \mathbb{R}^{P \times I}$ captures homogeneous dependencies and $\mathbf{B} \in \mathbb{R}^{Q \times I}$ captures inhomogeneous dependencies, where the sets P and Q index the rows of the matrices. Equations (15) need not be linearly independent but they are assumed to capture *all* existing affine dependencies, which means that

$$\operatorname{aff} \mathbf{Y} = \{\, \boldsymbol{\mu} \in \mathbb{R}^I \mid \mathbf{A}\boldsymbol{\mu} = \mathbf{0},\ \mathbf{B}\boldsymbol{\mu} = \mathbf{1} \,\} \tag{16}$$

If $\boldsymbol{\alpha} \in \mathbb{R}^P$ and $\boldsymbol{\beta} \in \mathbb{R}^Q$ are arbitrary *row* vectors and

$$\boldsymbol{\theta}' = \boldsymbol{\theta} - \boldsymbol{\alpha}\mathbf{A} - \boldsymbol{\beta}\mathbf{B} \tag{17}$$

then $\boldsymbol{\theta}'\mathbf{y} = \boldsymbol{\theta}\mathbf{y} - \boldsymbol{\beta}\mathbf{1}$ and $F(\boldsymbol{\theta}') = F(\boldsymbol{\theta}) - \boldsymbol{\beta}\mathbf{1}$. It follows that $q(\cdot\,|\,\boldsymbol{\theta}') = q(\cdot\,|\,\boldsymbol{\theta})$, i.e., the function $q(\cdot\,|\,\boldsymbol{\theta})$ is preserved by the transformations (17). This transformation is therefore called a *reparameterization* of the distribution. We further distinguish the subclass of transformations (17) given by $\boldsymbol{\theta}' = \boldsymbol{\theta} - \boldsymbol{\alpha}\mathbf{A}$ and call them *homogeneous* reparameterizations. The fact that $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are homogeneous reparameterization of each other is denoted by $\boldsymbol{\theta} \sim \boldsymbol{\theta}'$. While homogeneous reparameterizations preserve $\boldsymbol{\theta}\mathbf{y}$ and $F(\boldsymbol{\theta})$, general reparameterizations preserve only the difference[2] $\boldsymbol{\theta}\mathbf{y} - F(\boldsymbol{\theta})$, i.e., only distribution (10).

3.3. Mean parameters

The exponential family (10) naturally arises as follows: find a distribution $q(\mathbf{y})$ with maximum entropy and prescribed mean values $\boldsymbol{\mu} \in \mathbb{R}^I$ (a column vector with components μ_i) of the functions y_i, i.e., $\sum_{\mathbf{y} \in \mathbf{Y}} q(\mathbf{y})\mathbf{y} = \boldsymbol{\mu}$. This is why (1) or (10) is sometimes called the *maximal entropy model*. Solving this linearly constrained concave maximization task is easy and it reveals that $q(\mathbf{y})$ must have the form (10), where $\boldsymbol{\theta}$ appeared as Lagrange multipliers. Since entropy is strictly concave, $\boldsymbol{\mu}$ determines $q(\mathbf{y})$ uniquely. The numbers μ_i are called the *mean parameters* (or *moments*). Thus, any distribution from the family is uniquely given either by canonical parameters $\boldsymbol{\theta}$ or by mean parameters $\boldsymbol{\mu}$.

The *mean map* $\mathbf{m}\colon \mathbb{R}^I \to \mathbb{R}^I$ given by

$$\mathbf{m}(\boldsymbol{\theta}) = \sum_{\mathbf{y} \in \mathbf{Y}} q(\mathbf{y}\,|\,\boldsymbol{\theta})\mathbf{y} = \frac{\sum_{\mathbf{y} \in \mathbf{Y}} \mathbf{y} \exp \boldsymbol{\theta}\mathbf{y}}{\sum_{\mathbf{y} \in \mathbf{Y}} \exp \boldsymbol{\theta}\mathbf{y}} \tag{18}$$

assigns the mean parameters $\boldsymbol{\mu} = \mathbf{m}(\boldsymbol{\theta})$ to canonical parameters $\boldsymbol{\theta}$. The set of inverse elements is denoted by

$$\mathbf{m}^{-1}(\boldsymbol{\mu}) = \{\, \boldsymbol{\theta} \in \mathbb{R}^I \mid \mathbf{m}(\boldsymbol{\theta}) = \boldsymbol{\mu} \,\} \tag{19}$$

The ambiguity in solving the equation $\boldsymbol{\mu} = \mathbf{m}(\boldsymbol{\theta})$ is given exactly by reparametrizations; in other words, the set of all equivalents of $\boldsymbol{\theta}$ is equal to the set $\mathbf{m}^{-1}(\mathbf{m}(\boldsymbol{\theta}))$. Abusing notation, by $q(\mathbf{y}\,|\,\mathbf{m}^{-1}(\boldsymbol{\mu}))$ we denote the distribution from the family (uniquely) determined by $\boldsymbol{\mu}$.

The mean values of $\mathbf{y}$ over all possible distributions $q(\mathbf{y})$ form the *mean polytope*

$$\Big\{ \sum_{\mathbf{y} \in \mathbf{Y}} q(\mathbf{y})\mathbf{y} \;\Big|\; q(\mathbf{y}) \geq 0, \sum_{\mathbf{y} \in \mathbf{Y}} q(\mathbf{y}) = 1 \Big\} = \operatorname{conv} \mathbf{Y} \tag{20}$$

The equality in (20) is obvious by realizing that the scalars $q(\mathbf{y})$ play the role of the coefficients of convex combinations. Let us emphasize that $q(\mathbf{y})$ in (20) denotes arbitrary distributions, not necessarily from the exponential family. However, any vector $\boldsymbol{\mu}$ from the mean polytope can be realized also by a distribution *from* the exponential family, except the vectors on the boundary of the mean polytope, i.e., the range of the mapping $\mathbf{m}$ is the set $\operatorname{ri conv} \mathbf{Y}$. This follows from the fact that any $\boldsymbol{\mu} \in \operatorname{ri conv} \mathbf{Y}$ can be realized by infinitely many distributions $q(\mathbf{y})$ with the mean values $\boldsymbol{\mu}$, and restricting $q(\mathbf{y})$ to have the form (10) only picks the one with the maximum entropy. The boundary points are not covered because $\mathbf{m}(\boldsymbol{\theta})$ approaches the boundary of the mean polytope when and only when some components of $\boldsymbol{\theta}$ approach infinity [21],

$$\mathbf{m}(\boldsymbol{\theta}) \to \operatorname{rbd conv} \mathbf{Y} \quad \Longleftrightarrow \quad \|\boldsymbol{\theta}\| \to \infty \tag{21}$$

Note that vectors $\mathbf{y}$ and $\boldsymbol{\mu}$ are closely related: they both live in the mean polytope $\operatorname{conv} \mathbf{Y}$ but while $\boldsymbol{\mu}$ attain any value from (the inside of) $\operatorname{conv} \mathbf{Y}$, $\mathbf{y}$ is an element of the finite set $\mathbf{Y}$.

Using rule (12), one obtains the important equality

$$\frac{\mathrm{d}F(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} = \mathbf{m}(\boldsymbol{\theta}) \tag{22}$$

3.4. Entropy

One easily derives that the entropy of distribution $q(\mathbf{y}\,|\,\boldsymbol{\theta})$ is $F(\boldsymbol{\theta}) - \boldsymbol{\theta}\mathbf{m}(\boldsymbol{\theta})$. It follows that the function H defined by

$$\left.\begin{aligned} H(\boldsymbol{\mu}) &= F(\boldsymbol{\theta}) - \boldsymbol{\theta}\boldsymbol{\mu} \\ &\text{where } \boldsymbol{\theta} \text{ is arbitrary such that } \mathbf{m}(\boldsymbol{\theta}) = \boldsymbol{\mu} \end{aligned}\right\} \tag{23}$$

is the entropy of distribution $q(\mathbf{y}\,|\,\mathbf{m}^{-1}(\boldsymbol{\mu}))$ as a function of $\boldsymbol{\mu}$. The definition (23) is well-defined even for overcomplete parameterization because the number $F(\boldsymbol{\theta}) - \boldsymbol{\theta}\boldsymbol{\mu}$ is the same for all $\boldsymbol{\theta}$ satisfying $\mathbf{m}(\boldsymbol{\theta}) = \boldsymbol{\mu}$. The function H is positive and concave and its domain is $\operatorname{ri conv} \mathbf{Y}$.

Figure 1 shows plots of the log-partition function F, the mean map $\mathbf{m}$, the entropy function H, and the mean polytope $\operatorname{conv} \mathbf{Y}$ for a simple exponential family with $|I| = 2$ potential functions.

The relative entropy (KL-divergence) from a distribution $q(\mathbf{y}\,|\,\boldsymbol{\theta})$ to a (possibly different) distribution $q(\mathbf{y}\,|\,\mathbf{m}^{-1}(\boldsymbol{\mu}))$ is

$$D(\boldsymbol{\theta} \,\|\, \boldsymbol{\mu}) = F(\boldsymbol{\theta}) - H(\boldsymbol{\mu}) - \boldsymbol{\theta}\boldsymbol{\mu} \tag{24}$$

Any $\boldsymbol{\theta} \in \mathbb{R}^I$ and $\boldsymbol{\mu} \in \operatorname{ri conv} \mathbf{Y}$ satisfy

$$D(\boldsymbol{\theta} \,\|\, \boldsymbol{\mu}) \geq 0 \tag{25}$$

where the equality holds if and only if $\boldsymbol{\mu} = \mathbf{m}(\boldsymbol{\theta})$, i.e., if the distributions $q(\mathbf{y}\,|\,\boldsymbol{\theta})$ and $q(\mathbf{y}\,|\,\mathbf{m}^{-1}(\boldsymbol{\mu}))$ are equal.

Minimizing (24) allows us to express $F(\boldsymbol{\theta})$ and $H(\boldsymbol{\mu})$ of a single distribution in terms of each other as follows:

$$\sup_{\boldsymbol{\theta} \in \mathbb{R}^I} \; [\boldsymbol{\theta}\boldsymbol{\mu} - F(\boldsymbol{\theta})] = \begin{cases} -H(\boldsymbol{\mu}) & \text{if } \boldsymbol{\mu} \in \operatorname{ri conv} \mathbf{Y} \\ +\infty & \text{if } \boldsymbol{\mu} \notin \operatorname{conv} \mathbf{Y} \end{cases} \tag{26a}$$

[2] Thus, reparameterization can be understood in two slightly different meanings, according to whether only the distribution $q(\cdot\,|\,\boldsymbol{\theta})$ or also the unnormalized distribution $\boldsymbol{\theta}\mathbf{y}$ is preserved. In other texts, they often have only the first meaning, which we call here *homogeneous* reparameterization.

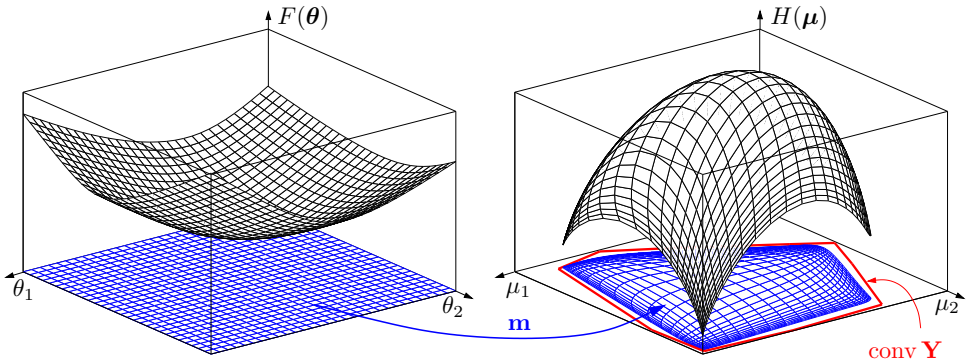


Figure 1. The plot of $F(\boldsymbol{\theta})$, $\mathbf{m}(\boldsymbol{\theta})$ and $H(\boldsymbol{\mu})$ for exponential family $(\mathbf{Y}, I)$, where $|\mathbf{Y}| = 256$, $I = \{1, 2\}$, and the numbers y_i were independently drawn from the normal distribution $\mathcal{N}[0, 1]$.

$$\max_{\boldsymbol{\mu}\in\operatorname{conv}\mathbf{Y}} [\boldsymbol{\theta}\boldsymbol{\mu} + H(\boldsymbol{\mu})] = F(\boldsymbol{\theta}) \tag{26b}$$

In (26a), we leave out the case $\boldsymbol{\mu} \in \operatorname{rbd}\mathbf{Y}$, which needs to be handled in a special way [21]. The convex minimization task (26a) attains its optimum at all vectors $\boldsymbol{\theta}$ satisfying $\boldsymbol{\mu} = \mathbf{m}(\boldsymbol{\theta})$. The concave maximization task (26b) attains its optimum at a single vector $\boldsymbol{\mu} = \mathbf{m}(\boldsymbol{\theta})$.

In fact, (26) shows that $F(\boldsymbol{\theta})$ and $-H(\boldsymbol{\mu})$ are related by convex conjugacy (Fenchel-Legendre transform) [3] and thus (25) can be alternatively interpreted as Fenchel's inequality.

For a minimal representation, similarly to (22) we have

$$-\frac{\mathrm{d}H(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}} = \mathbf{m}^{-1}(\boldsymbol{\mu}) \tag{27}$$

However, this is not valid for an overcomplete representations since H is defined on $\operatorname{ri}\operatorname{conv}\mathbf{Y} \subset \operatorname{aff}\mathbf{Y}$, where $\operatorname{aff}\mathbf{Y}$ is a strict subset of $\mathbb{R}^I$. Thus we cannot write simply $\mathrm{d}H(\boldsymbol{\mu})/\mathrm{d}\boldsymbol{\mu}$ but the derivative must be taken relative to affine space $\operatorname{aff}\mathbf{Y}$.

3.5. Generalized Iterative Scaling

Suppose we are given $\boldsymbol{\mu} \in \operatorname{ri}\operatorname{conv}\mathbf{Y}$ and want to find $\boldsymbol{\theta}$ such that $\boldsymbol{\mu} = \mathbf{m}(\boldsymbol{\theta})$. This is known as *moment fitting*. Of course, $\boldsymbol{\theta}$ cannot be found in closed-form – but it can be computed by a simple algorithm, the *Generalized Iterative Scaling* (GIS) [6, 5]. It assumes that $y_i \geq 0$ and $\sum_i y_i = 1$ for each $\mathbf{y} \in \mathbf{Y}$, which can be always achieved by affine transformations of vectors $\mathbf{y}$. Iterating the update

$$\theta_i \leftarrow \theta_i - \log m_i(\boldsymbol{\theta}) + \log \mu_i \tag{28}$$

converges to a state when $\boldsymbol{\mu} = \mathbf{m}(\boldsymbol{\theta})$. The algorithm monotonically increases the value of $\boldsymbol{\theta}\boldsymbol{\mu} - F(\boldsymbol{\theta})$, thus it can be understood as solving the optimization problem (26a). Despite its simplicity, proving convergence in argument is difficult [6, 5].

For the special case of undirected graphical models, the GIS algorithm reduces to the *Iterative Proportional Fitting* (IPF) procedure [7] and its analysis is much simpler.

3.6. Zero temperature limit

Let us look more formally at the zero temperature limit. Primarily, this means to investigate the behavior of the distribution $q(\mathbf{y}\,|\,\boldsymbol{\theta}/t)$ for the temperature t approaching zero. In turn, we can speak about the zero-temperature limit of not only $q(\mathbf{y}\,|\,\boldsymbol{\theta}/t)$ but also of the log-partition function F, the entropy H, and the mean map $\mathbf{m}$. Though only the limit $\bar{F}$ of F is directly useful for the inference tasks we consider, to give a complete picture we will discuss the limit also for p, H and $\mathbf{m}$. In the sequel, we denote the zero temperature limit of a quantity by placing a bar over it, e.g., F and $\bar{F}$.

The zero temperature limit of the log-partition function F (see §1) is given by

$$\bar{F}(\boldsymbol{\theta}) = \lim_{t\to 0+} tF(\boldsymbol{\theta}/t) \tag{29a}$$

$$= \max_{\mathbf{y}\in\mathbf{Y}} \boldsymbol{\theta}\mathbf{y} \tag{29b}$$

$$= \max_{\boldsymbol{\mu}\in\operatorname{conv}\mathbf{Y}} \boldsymbol{\theta}\boldsymbol{\mu} \tag{29c}$$

Equality (29a)=(29b) follows from the property (13) of the log-sum-exp operation. Equality (29b)=(29c) follows from the well-known fact that the optimum of a linear function in a polytope is always attained in at least one vertex.

Relation (29c) is important because it expresses $\bar{F}(\boldsymbol{\theta})$ as a linear programming over the mean polytope. Note, (29b) can be alternatively obtained as the limit of (26b), since

$$\lim_{t\to 0+} t[\boldsymbol{\theta}\boldsymbol{\mu}/t + H(\boldsymbol{\mu})] = \boldsymbol{\theta}\boldsymbol{\mu} \tag{30}$$

The zero temperature limit of distribution (10) trivially follows from (14):

$$\bar{q}(\mathbf{y}\,|\,\boldsymbol{\theta}) = \lim_{t\to 0+} q(\mathbf{y}\,|\,\boldsymbol{\theta}/t) = \begin{cases} |\mathbf{Y}^*(\boldsymbol{\theta})|^{-1} & \text{if } \mathbf{y} \in \mathbf{Y}^*(\boldsymbol{\theta}) \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{Y}^*(\boldsymbol{\theta}) = \operatorname{argmax}_{\mathbf{y}\in\mathbf{Y}} \boldsymbol{\theta}\mathbf{y}$ is the set of points with maximal probability. This proves the statement that $\bar{q}(\mathbf{y}\,|\,\boldsymbol{\theta})$ is nonzero if and only if $q(\mathbf{y}\,|\,\boldsymbol{\theta})$ is maximal.

Note, while the log-partition function $F(\boldsymbol{\theta})$ is the normalizing term of distribution $q(\mathbf{y}\,|\,\boldsymbol{\theta})$, nothing like this is true in the zero temperature limit: $\bar{F}(\boldsymbol{\theta})$ is *not* the normalizing term of $\bar{q}(\mathbf{y}\,|\,\boldsymbol{\theta})$.

The entropy of $\bar{q}(\mathbf{y}\,|\,\boldsymbol{\theta})$ is given by $-\log|\mathbf{Y}^*(\boldsymbol{\theta})|$. The mean function of $\bar{q}(\mathbf{y}\,|\,\boldsymbol{\theta})$ is

$$\bar{\mathbf{m}}(\boldsymbol{\theta}) = \lim_{t\to 0+} \mathbf{m}(\boldsymbol{\theta}/t) = |\mathbf{Y}^*(\boldsymbol{\theta})|^{-1} \sum_{\mathbf{y}\in\mathbf{Y}^*(\boldsymbol{\theta})} \mathbf{y} \tag{31}$$

i.e., $\bar{\mathbf{m}}(\boldsymbol{\theta})$ is the barycenter of the points $\mathbf{Y}^*(\boldsymbol{\theta})$ with maximum probability. The range of the function $\bar{\mathbf{m}}$ is the finite set, each element of which is the barycenter of the vertices of one face (of any dimension, i.e. including vertices) of $\operatorname{conv}\mathbf{Y}$.

4. Decomposition to subproblems

Here we develop the approach based on decomposition to subproblems, proposed in [23, 20]. In §4.2 we consider the non-zero temperature case, i.e., the upper bound on the log-partition function F. In §4.3, we consider the zero temperature case, i.e., the upper bound on $\bar{F}$. For both cases, we formulate the minimization of the upper bound, its Lagrange dual, and optimality conditions of this dual pair. In §4.4 we summarize these theoretical results and relate them to each other.

4.1. Notation for subfamilies of the exponential family

Further in §4, we will need to manipulate with subfamilies of the exponential family (10) obtained by reducing the set of its basis functions. That is, while family (10) is represented by $(\mathbf{Y}, I)$ where $\mathbf{Y} \subseteq \mathbb{R}^I$, we will consider families represented by $(\mathbf{Y}', I')$ given by

$$q'(\mathbf{y}' \,|\, \boldsymbol{\theta}') = \exp[\boldsymbol{\theta}'\mathbf{y}' - F(\boldsymbol{\theta}')] \tag{32}$$

where $\mathbf{y}' \in \mathbf{Y}'$, $I' \subseteq I$ and $\mathbf{Y}' \subseteq \mathbb{R}^{I'}$.

It will turn out convenient to express the log-partition, mean and entropy functions of the subfamily $(\mathbf{Y}', I')$ in terms of those of the family $(\mathbf{Y}, I)$. For that, we will represent the subset $I' \subseteq I$ by its characteristic vector $\boldsymbol{\delta} \in \{0,1\}^I$, such that

$$I' = \{\, i \in I \mid \delta_i = 1 \,\} \tag{33}$$

We further define operation $\boldsymbol{\delta}\bullet\mathbf{z}$ to be the componentwise product of two vectors, where the result is a column resp. row vector if $\mathbf{z}$ is column resp. row. This can be seen as the projection onto the dimensions given by non-zero componenets of $\boldsymbol{\delta}$. For a set $\mathbf{Z} \in \mathbb{R}^I$ we denote $\boldsymbol{\delta} \bullet \mathbf{Z} = \{\, \boldsymbol{\delta} \bullet \mathbf{z} \mid \mathbf{z} \in \mathbf{Z} \,\}$.

Now, the characteristics of subfamily $(\mathbf{Y}', I')$ can be expressed in terms of those of family $(\mathbf{Y}, I)$ as follows:

- The distribution is $q(\mathbf{y} \,|\, \boldsymbol{\delta} \bullet \boldsymbol{\theta})$.
- The log-partition function is $F(\boldsymbol{\delta} \bullet \boldsymbol{\theta})$.
- The mean polytope is $\boldsymbol{\delta} \bullet \operatorname{conv} \mathbf{Y} = \{\, \boldsymbol{\delta} \bullet \mathbf{y} \mid \mathbf{y} \in \operatorname{conv} \mathbf{Y} \,\}$, up to extra zero coordinates.
- The mean map is $\boldsymbol{\delta} \bullet \mathbf{m}(\boldsymbol{\delta} \bullet \boldsymbol{\theta})$, up to extra zero coordinates. Note that $\delta^i = 0$ does not imply that $m_i(\boldsymbol{\delta} \bullet \boldsymbol{\theta}) = 0$.
- Using (23), the entropy function is defined by

$$\left.\begin{aligned} &H_{\boldsymbol{\delta}}(\boldsymbol{\mu}) = F(\boldsymbol{\delta} \bullet \boldsymbol{\theta}) - (\boldsymbol{\delta} \bullet \boldsymbol{\theta})\boldsymbol{\mu} \\ &\quad \text{where } \boldsymbol{\theta} \text{ is arbitrary such that } \boldsymbol{\delta} \bullet \boldsymbol{\mu} = \boldsymbol{\delta} \bullet \mathbf{m}(\boldsymbol{\delta} \bullet \boldsymbol{\theta}) \end{aligned}\right\} \tag{34}$$

  where the domain of function $H_{\boldsymbol{\delta}}$ is the polyhedron

$$M_{\boldsymbol{\delta}} = \{\, \boldsymbol{\mu} \in \mathbb{R}^I \mid \boldsymbol{\delta} \bullet \boldsymbol{\mu} \in \boldsymbol{\delta} \bullet \operatorname{conv} \mathbf{Y} \,\} \tag{35}$$

  Note that $H_{\boldsymbol{\delta}}(\boldsymbol{\mu}) = H_{\boldsymbol{\delta}}(\boldsymbol{\delta} \bullet \boldsymbol{\mu})$ but $H_{\boldsymbol{\delta}}(\boldsymbol{\mu}) \neq H(\boldsymbol{\delta} \bullet \boldsymbol{\theta})$.

4.2. Non-zero temperature

Let $\{\, \boldsymbol{\theta}^s \in \mathbb{R}^I \mid s \in S \,\}$ be a collection of parameter vectors satisfying

$$\sum_{s \in S} \boldsymbol{\theta}^s \sim \boldsymbol{\theta} \tag{36}$$

Recall that $\sim$ denotes homogeneous reparameterization. Let scalars $\rho^s \geq 0$ be such that $\sum_{s \in S} \rho^s \geq 1$. Then

$$F(\boldsymbol{\theta}) = F\Big(\sum_{s \in S} \boldsymbol{\theta}^s\Big) \leq \sum_{s \in S} \rho^s F(\boldsymbol{\theta}^s / \rho^s) \tag{37}$$

If $\sum_{s \in S} \rho^s = 1$, inequality (37) follows from Jensen's inequality applied on the convex function F. If $\sum_{s \in S} \rho^s \geq 1$, inequality (37) follows from the property of the log-sum-exp operation given by (13). In fact, the inequality is looser for larger value of $\sum_{s \in S} \rho^s$.

The parameter vectors $\boldsymbol{\theta}^s$ are chosen such that the functions $F(\boldsymbol{\theta}^s)$ are tractable to compute, typically by setting most components of $\boldsymbol{\theta}^s$ to zero. The upper bound (37) is minimized over collections $\{\boldsymbol{\theta}^s\}$ such that (36) holds and the zero components of $\boldsymbol{\theta}^s$ are kept zero. Here, the free components of $\boldsymbol{\theta}^s$ are given by an indicator vector $\boldsymbol{\delta}^s \in \{0,1\}^I$ by requiring that $\boldsymbol{\theta}^s = \boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s$. These vectors form a collection

$$\Delta = \{\, \boldsymbol{\delta}^s \in \{0,1\}^I \mid s \in S \,\}$$

which is required to cover the whole set I, i.e.,

$$\max_{s \in S} \delta_i^s = 1 \qquad \forall i \in I \tag{38}$$

Now, the least upper bound is equal to

$$\min \Big\{ \sum_{s \in S} \rho^s F(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s / \rho^s) \;\Big|\; \boldsymbol{\theta}^s \in \mathbb{R}^I,\ \sum_{s \in S} \boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s \sim \boldsymbol{\theta} \Big\} \tag{39}$$

Now we form the dual to the convex minimization task (39).

Theorem 1. *The convex minimization task (39) and the concave maximization task*

$$\max \Big\{ \boldsymbol{\theta}\boldsymbol{\mu} + \sum_{s \in S} \rho^s H_{\boldsymbol{\delta}^s}(\boldsymbol{\mu}) \;\Big|\; \boldsymbol{\mu} \in \bigcap_{s \in S} M_{\boldsymbol{\delta}^s},\ \mathbf{A}\boldsymbol{\mu} = \mathbf{0} \Big\} \tag{40}$$

are related by strong Lagrange duality.

Proof. Note that (36) is satisfied if and only if $\sum_s \boldsymbol{\theta}^s = \boldsymbol{\theta} - \boldsymbol{\alpha}\mathbf{A}$ for some $\boldsymbol{\alpha}$. Denoting the Lagrange multipliers by $\boldsymbol{\mu}$, the Lagrangian of task (39) reads

$$L(\{\boldsymbol{\theta}^s\}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = (\boldsymbol{\theta} - \boldsymbol{\alpha}\mathbf{A})\boldsymbol{\mu} + \sum_{s \in S} [\rho^s F(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s / \rho^s) - (\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s)\boldsymbol{\mu}]$$

The dual reads $\max_{\boldsymbol{\mu} \in \mathbb{R}^I} L(\boldsymbol{\mu})$, where the Lagrange dual function $L(\boldsymbol{\mu}) = \inf_{\{\boldsymbol{\theta}^s\}, \boldsymbol{\alpha}} L(\{\boldsymbol{\theta}^s\}, \boldsymbol{\alpha}, \boldsymbol{\mu})$. If $\mathbf{A}\boldsymbol{\mu} \neq \mathbf{0}$ then $L(\boldsymbol{\mu}) = -\infty$, hence the constraint $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$ is needed. Let $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$. Then

$$L(\boldsymbol{\mu}) = \boldsymbol{\theta}\boldsymbol{\mu} + \sum_{s \in S} \inf_{\boldsymbol{\theta} \in \mathbb{R}^I} [\rho^s F(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta} / \rho^s) - (\boldsymbol{\delta}^s \bullet \boldsymbol{\theta})\boldsymbol{\mu}]$$

By (26a), we have

$$\inf_{\boldsymbol{\theta} \in \mathbb{R}^I} [\rho^s F(\boldsymbol{\delta} \bullet \boldsymbol{\theta} / \rho^s) - (\boldsymbol{\delta} \bullet \boldsymbol{\theta})\boldsymbol{\mu}] = \begin{cases} \rho^s H_{\boldsymbol{\delta}}(\boldsymbol{\mu}) & \text{if } \boldsymbol{\mu} \in \operatorname{ri} M_{\boldsymbol{\delta}} \\ -\infty & \text{if } \boldsymbol{\mu} \notin M_{\boldsymbol{\delta}} \end{cases}$$

Thus, the constraint $\boldsymbol{\mu} \in \bigcap_{\boldsymbol{\delta} \in \Delta} M_{\boldsymbol{\delta}}$ is needed. ∎

Next we formulate the conditions of joint optimality of the dual pair (39)+(40).

Theorem 2. *The following two statements are equivalent:*

- $\{\boldsymbol{\theta}^s\}$ *is optimal to (39) and* $\boldsymbol{\mu}$ *is optimal to (40).*

- $\{\boldsymbol{\theta}^s\}$ *and* $\boldsymbol{\mu}$ *satisfy*

$$\boldsymbol{\delta}^s \bullet \boldsymbol{\mu} = \boldsymbol{\delta}^s \bullet \mathbf{m}(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s / \rho^s) \quad \forall s \in S \tag{41a}$$

$$\sum_{s \in S} \boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s \sim \boldsymbol{\theta} \tag{41b}$$

$$\mathbf{A}\boldsymbol{\mu} = \mathbf{0} \tag{41c}$$

Proof. The conditions are easily obtained by setting the derivatives of $L(\{\boldsymbol{\theta}^s\}, \boldsymbol{\alpha}, \boldsymbol{\mu})$ w.r.t. $\boldsymbol{\theta}^s$, $\boldsymbol{\alpha}$ and $\boldsymbol{\mu}$ to zero, using (22) and some simple manipulations. ∎

4.3. Zero temperature

Similarly as (37), we can write the upper bound on $\bar{F}$:

$$\bar{F}(\boldsymbol{\theta}) = \bar{F}\Big(\sum_{s \in S} \boldsymbol{\theta}^s\Big) \le \sum_{s \in S} \bar{F}(\boldsymbol{\theta}^s) \tag{42}$$

Since $\bar{F}(\boldsymbol{\theta}/\rho) = \bar{F}(\boldsymbol{\theta})/\rho$, the scalar ρ in (37) cancels out.

Let us write the dual pair expressing the least upper bound (42), analogical to (39)+(40). Taking the zero temperature limit of the primal (39) results in $F(\boldsymbol{\theta})$ being replaced with $\bar{F}(\boldsymbol{\theta})$, as shown by (29):

$$\min\Big\{ \sum_{s \in S} \bar{F}(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s) \;\Big|\; \boldsymbol{\theta}^s \in \mathbb{R}^I,\ \sum_{s \in S} \boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s \sim \boldsymbol{\theta} \Big\} \tag{43}$$

Taking the limit of the dual (40) results in the entropy term vanishing, as shown by (30):

$$\max\Big\{ \boldsymbol{\theta}\boldsymbol{\mu} \;\Big|\; \boldsymbol{\mu} \in \bigcap_{\boldsymbol{\delta} \in \Delta} M_{\boldsymbol{\delta}},\ \mathbf{A}\boldsymbol{\mu} = \mathbf{0} \Big\} \tag{44}$$

We give a formal proof that the tasks (43) and (44) are really dual. Given the above arguments the proof may seem unnecessary – but we will refer to this proof later in Theorem 4.

Theorem 3. *The tasks (43) and (44) are related by linear programming duality.*

Proof. Consider the following simple pair of linear programs:

$$h \to \min \quad \boldsymbol{\theta} \sum_{\mathbf{y} \in \mathbf{Y}} q(\mathbf{y})\mathbf{y} \to \max \tag{45a}$$

$$\boldsymbol{\theta}\mathbf{y} \le h \qquad q(\mathbf{y}) \ge 0 \qquad \forall \mathbf{y} \in \mathbf{Y} \tag{45b}$$

$$h \in \mathbb{R} \qquad \sum_{\mathbf{y} \in \mathbf{Y}} q(\mathbf{y}) = 1 \tag{45c}$$

It is quite obvious that the linear programs (45) are dual to each other and their common optimum equals $\bar{F}(\boldsymbol{\theta})$. The constraints in the pair are written such that a variable and its Lagrange multiplier is on the same line.

In a similar way, the tasks (43) and (44) can be respectively written as the following linear programs:

$$\sum_{s \in S} h^s \to \min \qquad \boldsymbol{\theta}\boldsymbol{\mu} \to \max \tag{46a}$$

$$\sum_{s \in S} \boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s + \boldsymbol{\alpha}\mathbf{A} = \boldsymbol{\theta} \qquad \boldsymbol{\mu} \in \mathbb{R}^I \tag{46b}$$

$$(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s)\mathbf{y} \le h^s \qquad q^s(\mathbf{y}) \ge 0 \qquad \forall s, \mathbf{y} \tag{46c}$$

$$\boldsymbol{\theta}^s \in \mathbb{R}^I \quad \boldsymbol{\delta}^s \bullet \sum_{\mathbf{y} \in \mathbf{Y}} q^s(\mathbf{y})\mathbf{y} = \boldsymbol{\delta}^s \bullet \boldsymbol{\mu} \quad \forall s \tag{46d}$$

$$h^s \in \mathbb{R} \qquad \sum_{\mathbf{y} \in \mathbf{Y}} q^s(\mathbf{y}) = 1 \qquad \forall s \tag{46e}$$

$$\boldsymbol{\alpha} \in \mathbb{R}^P \qquad \mathbf{A}\boldsymbol{\mu} = 0 \tag{46f}$$

It can be verified that either program can be constructed from the other one by LP duality. ∎

Next we give optimality conditions, analogical to (41).

Theorem 4. *The following two statements are equivalent:*

- $\{\boldsymbol{\theta}^s\}$ *is optimal to (43) and* $\boldsymbol{\mu}$ *is optimal to (44).*
- $\{\boldsymbol{\theta}^s\}$ *and* $\boldsymbol{\mu}$ *satisfy*

$$\boldsymbol{\delta}^s \bullet \boldsymbol{\mu} \in \boldsymbol{\delta}^s \bullet \operatorname*{argmax}_{\boldsymbol{\mu} \in \operatorname{conv} \mathbf{Y}} (\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s)\boldsymbol{\mu} \quad \forall s \in S \tag{47a}$$

$$\sum_{s \in S} \boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s \sim \boldsymbol{\theta} \tag{47b}$$

$$\mathbf{A}\boldsymbol{\mu} = \mathbf{0} \tag{47c}$$

Proof. We shall show that (47a) are the complementary slackness conditions for the LP pair (46).

First, we can eliminate the variables h^s because at optimum, we have $h^s = \bar{F}(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s) = \max_{\mathbf{y} \in \mathbf{Y}} (\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s)\mathbf{y}$. By complementary slackness, the two inequalities on line (46c) are never simultaneously strict, i.e., for all s and $\mathbf{y}$ we have

$$\big[(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s)\mathbf{y} - \max_{\mathbf{y} \in \mathbf{Y}} (\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s)\mathbf{y}\big] q^s(\mathbf{y}) = 0 \tag{48}$$

Now, realize that the following two sets are equal:

$$\operatorname*{argmax}_{\boldsymbol{\mu} \in \operatorname{conv} \mathbf{Y}} \boldsymbol{\theta}\boldsymbol{\mu} = \Big\{ \sum_{\mathbf{y} \in \mathbf{Y}} q(\mathbf{y})\mathbf{y} \;\Big|\; q(\mathbf{y}) \ge 0, \sum_{\mathbf{y} \in \mathbf{Y}} q(\mathbf{y}) = 1,\ \big[\boldsymbol{\theta}\mathbf{y} - \max_{\mathbf{y} \in \mathbf{Y}} \boldsymbol{\theta}\mathbf{y}\big] q(\mathbf{y}) = 0 \Big\}$$

The last constraint in the right-hand set is the same as (48), up to projections onto subspaces $\boldsymbol{\delta}^s$. Taking into account (46d), this shows that $\boldsymbol{\mu}$ satisfies (47a). ∎

4.4. Summary of the section

The theoretical results obtained so far in §4 may look complex. Here we interpret them and relate them to each other.

In the sequel, we set $\rho^s = 1$ for simplicity. Of course, inequality (37) remains satisfied for this case. However, if we are interested in upper bounds of the log-partition function F in the sense of [22], we need to keep in mind that inequality (37) is tightest for $\sum_{s \in S} \rho^s = 1$ and (very) loose for $\rho^s = 1$.

4.4.1 Primal tasks

Let us first recall the convex tasks (39) and (43) of minimizing the upper bound on F and $\bar{F}$:

$$\min\Big\{ \sum_{s \in S} F(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s) \;\Big|\; \boldsymbol{\theta}^s \in \mathbb{R}^I,\ \sum_{s \in S} \boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s \sim \boldsymbol{\theta} \Big\} \tag{49a}$$

$$\min\Big\{ \sum_{s \in S} \bar{F}(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s) \;\Big|\; \boldsymbol{\theta}^s \in \mathbb{R}^I,\ \sum_{s \in S} \boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s \sim \boldsymbol{\theta} \Big\} \tag{49b}$$

4.4.2 Dual tasks

First, recall formulas (26b) and (29b), showing that the true functions F and $\bar{F}$ can be expressed as optimizations over the

mean polytope:

$$F(\boldsymbol{\theta}) = \max_{\boldsymbol{\mu} \in \operatorname{conv} \mathbf{Y}} [\boldsymbol{\theta}\boldsymbol{\mu} + H(\boldsymbol{\mu})] \tag{50a}$$

$$\bar{F}(\boldsymbol{\theta}) = \max_{\boldsymbol{\mu} \in \operatorname{conv} \mathbf{Y}} \boldsymbol{\theta}\boldsymbol{\mu} \tag{50b}$$

Now, notice that the duals (40) and (44) of (49) can be written in a similar form:

$$F(\boldsymbol{\theta}) \leq \max_{\boldsymbol{\mu} \in M_\Delta} \Big[\boldsymbol{\theta}\boldsymbol{\mu} + \sum_{\boldsymbol{\delta} \in \Delta} H_{\boldsymbol{\delta}}(\boldsymbol{\mu})\Big] \tag{51a}$$

$$\bar{F}(\boldsymbol{\theta}) \leq \max_{\boldsymbol{\mu} \in M_\Delta} \boldsymbol{\theta}\boldsymbol{\mu} \tag{51b}$$

where the polyhedron M_Δ is given by

$$M_\Delta = \{\, \boldsymbol{\mu} \in \mathbb{R}^I \mid \mathbf{A}\boldsymbol{\mu} = \mathbf{0} \,\} \cap \bigcap_{\boldsymbol{\delta} \in \Delta} M_{\boldsymbol{\delta}} \tag{52}$$

where (see (35))

$$M_{\boldsymbol{\delta}} = \{\, \boldsymbol{\mu} \in \mathbb{R}^I \mid \boldsymbol{\delta} \bullet \boldsymbol{\mu} \in \boldsymbol{\delta} \bullet \operatorname{conv} \mathbf{Y} \,\} \tag{53}$$

This shows the difference between the variational formulation (50) for the true F and $\bar{F}$ and for their upper bounds (51): the latter is obtained from the former by replacing the exact entropy function H with the *sum of entropies $H_{\boldsymbol{\delta}}$ of the subproblems* and the mean polytope $\operatorname{conv} \mathbf{Y}$ with M_Δ.

The set (52) is the intersection of polyhedra $M_{\boldsymbol{\delta}}$ and the linear space $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$. The intersection $\bigcap_{\boldsymbol{\delta} \in \Delta} M_{\boldsymbol{\delta}}$ is *the largest polyhedron that has the same projection onto each subspace $\boldsymbol{\delta} \in \Delta$ as the mean polytope* $\operatorname{conv} \mathbf{Y}$ (Figure 2 visualizes this construction for $|I| = 3$). It follows that M_Δ is an *outer bound of the mean polytope* $\operatorname{conv} \mathbf{Y}$.

Without assuming (38), M_Δ can be an unbounded polyhedron. However, if (38) holds, M_Δ is bounded (i.e., a polytope) because its projections $\boldsymbol{\delta} \bullet \operatorname{conv} \mathbf{Y}$ are bounded and each coordinate is covered by at least one subproblem.

While $\boldsymbol{\mu} \in \operatorname{conv} \mathbf{Y}$ can be obtained as mean parameters of the exponential family (10) for some $\boldsymbol{\theta}$, this may not be true for $\boldsymbol{\mu} \in M_\Delta$. Since $\operatorname{conv} \mathbf{Y} \subseteq M_\Delta$, a vector $\boldsymbol{\mu} \in M_\Delta \setminus \operatorname{conv} \mathbf{Y}$ does not correspond to mean parameters of family (10) for any $\boldsymbol{\theta}$. Therefore, $\boldsymbol{\mu} \in M_\Delta$ can be called *pseudo-mean parameters* and M_Δ can be called the *pseudo-mean polytope*.

4.4.3 Optimality conditions

Let us now focus on conditions (41), stating when the upper bound on F is optimal. Condition (41a) can be formulated in a more transparent way as the set of equations

$$\mu_i = m_i(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s) \qquad \forall s \in S,\ i \in I,\ \delta_i^s = 1 \tag{54}$$

This shows that optimality of the pair (39)+(40) is characterized by the fact that *the mean parameters of overlapping subproblems are equal*.

In the zero temperature limit, the situation becomes different. The conditions for joint optimality of the pair (43)+(44) are given by (47). Although expression (47a) may look complex, it has a clear interpretation. The set $\operatorname{argmax}_{\boldsymbol{\mu} \in \operatorname{conv} \mathbf{Y}} \boldsymbol{\theta}\boldsymbol{\mu}$ is the optimal face of the mean polytope with respect to the linear function $\boldsymbol{\theta}\boldsymbol{\mu}$. Thus, the set on the right-hand side of (47a) is the optimal face of the mean polytope of subproblem s, up to extra zero coordinates. Multiplying both sides of (47a) by $\boldsymbol{\delta}^s$ means that μ_i is arbitrary whenever $\delta_i^s = 0$. Hence, (47a) requires that *the optimal faces of the mean polytopes of the subproblems (pulled back to the original space $\mathbb{R}^I$) have a non-empty intersection* (see Figure 3).

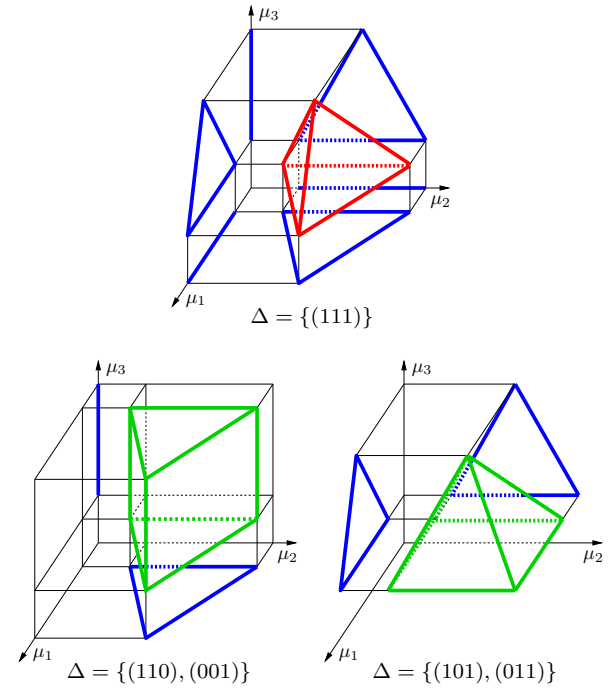


Figure 2. The top subfigure shows an example of the mean polytope $\operatorname{conv} \mathbf{Y}$ for $I = \{1, 2, 3\}$ (in red) and its projections onto dimensions $\boldsymbol{\delta} = (100), (010), (001), (011), (101), (110)$ (in blue). The bottom subfigures show the polyhedron $\bigcap_{\boldsymbol{\delta} \in \Delta} M_{\boldsymbol{\delta}}$ (in green) for two different collections Δ.

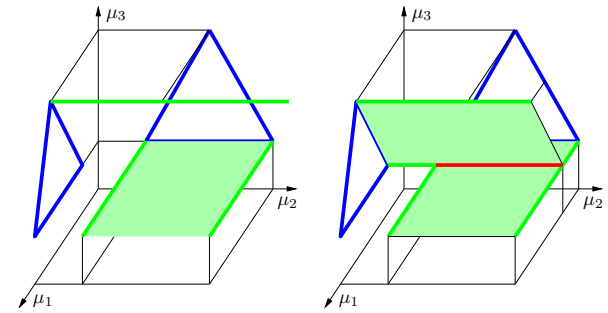


Figure 3. Intersection of the optimal faces of the subproblem polyhedra $M_{\boldsymbol{\delta}}$ (in blue) shown in Figure 2 for $\Delta = \{(101), (011)\}$. In the left subfigure, the optimal faces pulled back to the original space (in green) do not intersect. In the right-hand subfigure, they intersect in the 1-dimensional intersection (in red).

In fact, we can write the optimality condition for both non-zero and zero temperature as a single formula:

$$\boldsymbol{\delta}^s \bullet \boldsymbol{\mu} \in \boldsymbol{\delta}^s \bullet \operatorname*{argmax}_{\boldsymbol{\mu} \in \operatorname{conv} \mathbf{Y}} [(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s)\boldsymbol{\mu} + t H_{\boldsymbol{\delta}^s}(\boldsymbol{\mu})] \quad \forall s \in S \tag{55}$$

For $t = 1$, the right-hand set has only a single element (because entropy is strictly concave) and thus (55) reduces to (54). For $t \to 0+$, the right-hand set has in general an infinite number of solutions and (55) reduces to (47a).

Note that conditions (47a) can be seen as a generalization

of the subproblem agreement (8), given by

$$\mathbf{y} \in \operatorname*{argmax}_{\mathbf{y}\in\mathbf{Y}} \boldsymbol{\theta}^s \mathbf{y} \qquad \forall s \in S \tag{56}$$

4.4.4 When overcompleteness can be ignored

Sometimes, expression $\sum_s \boldsymbol{\theta}^s \sim \boldsymbol{\theta}$ in (39) can be replaced with $\sum_s \boldsymbol{\theta}^s = \boldsymbol{\theta}$ without affecting the optimum. This would be of course convenient: it would simplify many expressions in §4, removing the symbol $\sim$ and constraint $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$. Theorem 5 specifies when this is possible[3]. We shall see in §7.2 that for MRFs, the assumption of the theorem has a clear interpretation.

Theorem 5. *Let every $\boldsymbol{\mu} \in \bigcap_{\boldsymbol{\delta}\in\Delta} M_{\boldsymbol{\delta}}$ satisfy $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$ (i.e., the polyhedron $\bigcap_{\boldsymbol{\delta}\in\Delta} M_{\boldsymbol{\delta}}$ is contained in the linear subspace $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$). Then replacing $\sim$ with $=$ does not change the optimal value of (39) and (43), and removing the constraint $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$ does not change the optimal value of (40) and (44).*

Proof. The constraint $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$ corresponds via duality to Lagrange multiplier $\boldsymbol{\alpha}$. When $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$ is implied by other constraints, it can be removed along with its multiplier. ∎

5. Coordinate descent

In this section, we present a (block-)coordinate descent algorithm to minimize (or at least decrease) the upper bound on F and $\bar{F}$, i.e., to solve the optimization tasks (39) and (43). This will be done for the *general form* of the distribution (10), where $(\mathbf{Y}, I)$ are not constrained to define a graphical model. Our algorithm has two sources of inspiration:

- Generalized Iterative Scaling (see §3.5), which fits canonical parameters $\boldsymbol{\theta}$ into given mean parameters $\boldsymbol{\mu}$.
- Max-sum diffusion [12, 28, 30, 29] and similar algorithms [10, 9], based on averaging (max-)marginals in MRFs.

We have not been able to formulate coordinate descent for the completely arbitrary exponential family (10). We have achieved the following:

- For non-zero temperature (§5.1), we give an algorithm that strictly and monotonically decreases the upper bound under assumption that $0 \le y_i \le 1$. This is not a restriction since it can be always achieved by an affine transformation of vectors $\mathbf{y}$.
- For zero temperature (§5.2), we give an algorithm that monotonically (but not strictly) decreases the upper bound under assumption that $y_i \in \{0, 1\}$. This algorithm is a straightforward modification of the algorithm for non-zero temperature.

Throughout §5, it is supposed that the assumptions of Theorem 5 hold, hence $\sim$ can be replaced with $=$ and the constraint $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$ can be discarded in (39), (40), (41), (43), (44), (47). We believe it would not be hard to relax this assumption.

[3] The problem '$\sim$ vs. $=$' has been discussed also in [10].

5.1. Non-zero temperature

Recall (§3.5) that the equation $\boldsymbol{\mu} = \mathbf{m}(\boldsymbol{\theta})$ can be solved by the GIS algorithm, which is equivalent to solving the optimization task (26a). In GIS, we iterate an update $\theta_i \leftarrow \theta_i + \xi$, where ξ is chosen such that

(a) the objective $\boldsymbol{\theta}\boldsymbol{\mu} - F(\boldsymbol{\theta})$ improves,

(b) $\xi = 0$ if and only if $\mu_i = m_i(\boldsymbol{\theta})$.

The choice $\xi = \log \mu_i - \log m_i(\boldsymbol{\theta})$ in (28) satisfies (a) and (b).

To minimize the upper bound on the log-partition function, we are in a similar (though more complex) situation. We need to solve the optimization task (39), which is equivalent to solving the equation system

$$\mu_i = m_i(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s) \qquad \forall i \in I,\ s \in S_i \tag{57a}$$

$$\sum_{s\in S_i} \theta_i^s = \theta_i \qquad \forall i \in I \tag{57b}$$

for $\{\boldsymbol{\theta}^s\}$ and $\boldsymbol{\mu}$, where we denoted by $S_i = \{\, s \in S \mid \delta_i^s = 1 \,\}$ the subset of the subproblems overlapping at $i \in I$. While in GIS we fit $\boldsymbol{\theta}$ into given mean parameters $\boldsymbol{\mu}$, here we want to find $\boldsymbol{\theta}$ satisfying (57b) such that the mean parameters of overlapping subproblems coincide. To do it, we iterate an update in which we pick a single $i \in I$ and change the block of coordinates $\{\, \theta_i^s \mid s \in S_i \,\}$ as

$$\theta_i^s \leftarrow \theta_i^s + \xi^s \qquad \forall s \in S_i \tag{58}$$

where the correction terms $\{\, \xi^s \mid s \in S_i \,\}$ are such that

(a') the objective $\sum_{s\in S} F(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s)$ improves,

(b') $\xi^s = 0$ if and only if the mean parameters $\{\, m_i(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s) \mid s \in S_i \,\}$ are the same,

(c') $\sum_{s\in S_i} \xi_s = 0$, hence condition (57b) is kept satisfied.

We show how to choose such $\{\xi^s\}$. Let

$$\xi^s = \left[\frac{1}{|S_i|} \sum_{s'\in S_i} n_i(\boldsymbol{\delta}^{s'} \bullet \boldsymbol{\theta}^{s'})\right] - n_i(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s) \tag{59}$$

where $n_i(\boldsymbol{\theta}) = \log m_i(\boldsymbol{\theta})$.

Quite obviously, $\{\xi^s\}$ defined by (59) satisfy conditions (b') and (c') above. It is much less trivial to prove that they satisfy also condition (a'). It is given by Theorem 6, where we assume $S_i = S$ without loss of generality. It is easy to verify that $\{\xi^s\}$ satisfy the assumption (60) of the theorem.

Theorem 6. *Let $\theta_i^s \in \mathbb{R}$ and $0 \le y_i \le 1$. Pick a single $i \in I$. Let $\{\, \xi^s \mid s \in S \,\}$ be any numbers satisfying*

$$\bigoplus_{s\in S} [n_i(\boldsymbol{\theta}^s) + \xi^s] < \bigoplus_{s\in S} n_i(\boldsymbol{\theta}^s) \tag{60}$$

Then updating the parameters $\{\, \theta_i^s \mid s \in S \,\}$ as $\theta_i^s \leftarrow \theta_i^s + \xi^s$ strictly decreases the expression $\sum_{s\in S} F(\boldsymbol{\theta}^s)$.

Proof. In appendix A. ∎

This update (58) is done *sequentially*, i.e., the function $\mathbf{n}$ is newly calculated for each new i, such as shown in Algorithm 1. The update (58) need not find the global minimum of the upper bound over the block of coordinates $\{\, \theta_i^s \mid s \in S_i \,\}$,

Algorithm 1 (coordinate descent)

repeat
  for $i \in I$ **do**
    for $s \in S_i$ **do** $\nu_i^s \leftarrow n_i(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s)$ **end for**
    $\nu_i \leftarrow |S_i|^{-1} \sum_{s \in S_i} \nu_i^s$
    for $s \in S_i$ **do** $\theta_i^s \leftarrow \theta_i^s - \nu_i^s + \nu_i$ **end for**
  end for
until convergence

it only ensures a strict improvement whenever the optimality conditions are not satisfied.

Of course, properties (a')+(b')+(c') may not imply convergence to the solution, they only guarantee monotonic and strict improvement of the bound.

5.2. Zero temperature

Let us now focus on minimizing the upper bound on $\bar{F}$, i.e., on solving the optimization problem (43). Coordinate descent from §5.1 can be very naturally modified for (43): simply *replace each occurrence of the log-sum-exp operation $\oplus$ with the ordinary maximum*, like in tropical algebra [14, 8]. To do it, we first express the functions $n_i(\boldsymbol{\theta})$ used in (59) in a different form, containing the operation $\oplus$. Obviously, (18) can be alternatively written as

$$\log m_i(\boldsymbol{\theta}) = n_i(\boldsymbol{\theta}) = F_i(\boldsymbol{\theta}) - F(\boldsymbol{\theta}) \tag{61}$$

where

$$F_i(\boldsymbol{\theta}) = \bigoplus_{\mathbf{y} \in \mathbf{Y}} [\boldsymbol{\theta}\mathbf{y} + \log y_i] \tag{62}$$

Now, let $\{\xi^s\}$ in (58) be given as

$$\xi^s = \left[\frac{1}{|S_i|} \sum_{s' \in S_i} \bar{n}_i(\boldsymbol{\delta}^{s'} \bullet \boldsymbol{\theta}^{s'}) \right] - \bar{n}_i(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s) \tag{63}$$

where the functions $\bar{n}_i$ are defined as n_i where $\oplus$ is replaced with $\max$:

$$\bar{n}_i(\boldsymbol{\theta}) = \bar{F}_i(\boldsymbol{\theta}) - \bar{F}(\boldsymbol{\theta}) \tag{64a}$$

$$\bar{F}_i(\boldsymbol{\theta}) = \max_{\mathbf{y} \in \mathbf{Y}} [\boldsymbol{\theta}\mathbf{y} + \log y_i] \tag{64b}$$

Thus, Algorithm 1 can be used after replacing n_i with $\bar{n}_i$.

Then, Theorem 6 holds which is very similar than Theorem 6 but weaker in two ways. First, vectors $\mathbf{y}$ are required to satisfy $y_i \in \{0, 1\}$ rather than only $0 \leq y_i \leq 1$. Note that in this case (since $\log 0 = -\infty$), expression (64b) can be more conveniently written as

$$\bar{F}_i(\boldsymbol{\theta}) = \max_{\mathbf{y} \in \mathbf{Y} | y_i = 1} \boldsymbol{\theta}\mathbf{y} = \theta_i + \max_{\mathbf{y} \in \mathbf{Y} | y_i = 1} \sum_{j \in I \setminus i} \theta_j y_j \tag{65}$$

where the second equality is easy to verify. Second, the upper bound may not always strictly decrease – sometimes it decreases and sometimes it remains unchanged.

Theorem 7. *Let $\theta_i^s \in \mathbb{R}$ and $y_i \in \{0, 1\}$. Pick a single $i \in I$. Let $\{\, \xi^s \mid s \in S \,\}$ be any numbers satisfying*

$$\max_{s \in S} [\bar{n}_i(\boldsymbol{\theta}^s) + \xi^s] < \max_{s \in S} \bar{n}_i(\boldsymbol{\theta}^s) \tag{66}$$

Then updating the parameters $\{\, \theta_i^s \mid s \in S \,\}$ as $\theta_i^s \leftarrow \theta_i^s + \xi^s$ does not increase the expression $\sum_{s \in S} \bar{F}(\boldsymbol{\theta}^s)$.

Proof. In appendix A. ∎

Theorem 7 is obviously not sufficient to guarantee a strict improvement of the bound, it only states that the bound does not become worse. To guarantee strict improvement, we would need a stronger theorem. In analogy with MRFs, such a theorem could look as follows. There exists a property P of collection $\{\boldsymbol{\theta}^s\}$ such that if P is not satisfied then after a finite number of updates the upper bound strictly improves. In MRFs, P is *weak tree agreement* [10], more widely known as *arc consistency* [28]. We did not find such a property for general exponential families and we do not know whether it exists.

Note the following important subtlety. The fixed point of the non-zero temperature version of Algorithm 1 is characterized by equality of the mean parameters of overlapping subproblems given by (57), i.e.

$$\forall i \in I \; \exists \mu_i \; \forall s \in S_i : \quad \mu_i = m_i(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s) \tag{67}$$

After substituting (61) we can write (67) equivalently in the logarithmic domain as

$$\forall i \in I \; \exists \nu_i \; \forall s \in S_i : \quad \nu_i = n_i(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s) \tag{68}$$

The fixed point of the zero temperature version of Algorithm 1 is characterized by the condition

$$\forall i \in I \; \exists \bar{\nu}_i \; \forall s \in S_i : \quad \bar{\nu}_i = \bar{n}_i(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s) \tag{69}$$

While for non-zero temperature we have two equivalent conditions (68) and (67), for zero temperature we have only condition (69) but there is no analogy of condition (67). This is because there exists no zero-temperature version of equality (61) (see §3.6): the zero temperature limit $\bar{\mathbf{m}}$ of the mean map given by (31) is not related to $\bar{\mathbf{n}}$ in any simple way.

The fixed point (69), obtained by mechanically replacing $\oplus$ with $\max$, is different from condition (47), necessary and sufficient for optimality of the upper bound under zero temperature. Thes two conditions are incomparable, (69) is neither stronger nor weaker than (47). It follows that the algorithm in general may not find the global minimum of the upper bound. This is not surprising as we are solving the minimization task (43) which is convex but *nonsmooth*. It is well-known [2] and rather obvious (see Figure 4) that *(block-)coordinate descent does not find the global minimum of a nonsmooth convex function* in general. It finds a point that is globally minimal for each coordinate separately but may not be for all coordinates simultaneously. This can be seen as a 'local minimum', where locality is defined with respect to the coordinate moves.

We carried out experiments in which we chose small sets $\mathbf{Y}$ and I (where $0 \leq y_i \leq 1$ for non-zero temperature and $y_i \in \{0, 1\}$ for zero temperature) and random $\boldsymbol{\theta}$. Subproblem collection Δ was generated randomly, subject to the covering condition (38). We observed that both the non-zero and zero temperature version of Algorithm 1 always monotonically decreased the upper bound and converged to a state when the fixed conditions were satisfied.

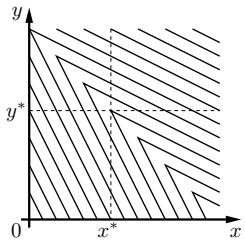


Figure 4. The figure shows the contours of a non-smooth convex function, $f(x, y)$. Coordinate descent did not find the global minimum: point (x^*, y^*) is globally minimal for each coordinate separately but not for both simultaneously.

5.3. A note on interior point algorithm

The fact that the coordinate descent algorithm does may not find the global minimum of the upper bound can be alleviated by running the algorithm for gradually decreasing sequence of non-zero temperatures, converging to zero. This *smoothing method* to avoid local minima of a convex non-smooth function has been discussed elsewhere [30, §7.4], [25, 9, 16].

From the point of view of the dual tasks (51), these algorithms are *interior point algorithms* because as the temperature decreases, the current optimal solution $\boldsymbol{\mu}$ moves from the interior of the polytope M_Δ to its boundary. Here, the function $\sum_{\boldsymbol{\delta}\in\Delta} H_{\boldsymbol{\delta}}(\boldsymbol{\mu})$ in problem (40) plays the role of the *barrier function*. A typical barrier functions approach infinity when the parameter approaches the boundary of the feasible set. In contrast, it follows from (27) and (21) that when $\boldsymbol{\mu}$ approaches the boundary of $\operatorname{conv}\mathbf{Y}$, our barrier function remains finite but the magnitude of its *derivative* approaches infinity.

6. Cutting plane algorithm

By including more and more complex (but tractable, such as cycles or graphs with limited treewidth in case of MRFs) subproblems into collection Δ, one obtains a hierarchy of increasingly tighter upper bounds. Moreover, not all subproblems need be present in Δ from the beginning – they can be added incrementally. We can start with Δ containing some simple subproblems (such as trees in case of MRFs) and incrementally add more complex ones (such as cycles or problems with small treewidth).

6.1. Zero temperature

Let us first consider the cutting plane strategy for zero temperature, i.e., to upper-bound $\bar{F}$. If the least upper bound given by a current collection Δ turns out to be loose, we can add a suitable subproblem $\boldsymbol{\delta} \notin \Delta$ to Δ. If we initialize the canonical parameters of this added subproblem to zero, the current upper bound $\sum_s \bar{F}(\boldsymbol{\delta}^s \bullet \boldsymbol{\theta}^s)$ on the mode $\bar{F}$ does not change – this is because $\bar{F}(\boldsymbol{\theta}) = 0$ for $\boldsymbol{\theta} = \mathbf{0}$. Subsequent iterations of Algorithm 1 will either improve or preserve the upper bound.

In terms of pseudomarginals (i.e., of our dual tasks (51)), an improvement would correspond to cutting off a part of the pseudo-mean polytope M_Δ that is not in the mean polytope $\operatorname{conv}\mathbf{Y}$. This strategy is well-known in integer programming as the *cutting plane algorithm*. Indentifying a subproblem whose addition would ensure a bound improvement then corresponds to the *separation problem*. The separation problem can be easily formulated as follows: given a current optimal vector $\boldsymbol{\mu}$ of pseudomarginals, find $\boldsymbol{\delta}$ such that $\boldsymbol{\delta} \bullet \boldsymbol{\mu} \notin M_{\boldsymbol{\delta}}$.

We demonstrated this algorithm for the special case of MRFs in [26].

6.2. Non-zero temperature

There seems to be an obstacle to using this form of the cutting plane strategy for non-zero temerature, i.e., to upper-bound F. After initializing the canonical parameters of an added subproblem to zero, the current upper bound on $\bar{F}$ does not change because $\bar{F}(\mathbf{0}) = 0$. However, the upper bound on F *increases* because $F(\mathbf{0}) = \log|\mathbf{Y}| > 0$. This undesirable effect might be reduced by maintaining $\sum_s \rho^s = 1$ before and after adding the subproblem (note that for zero temperature, ρ^s cancel out). We do not know whether the effect can be removed entirely by some straightforward modification or whether it is an inherent obstacle to using the cutting plane strategy to upper-bound the partition function.

One might conjecture that the effect would dissapear if the upper bound on F (given by (37)) was minimized not only over $\{\boldsymbol{\theta}^s\}$ but also over $\{\rho^s\}$. The evidence (not proof) why this could be so is as follows: if we add the subproblem $\boldsymbol{\delta}^r = \mathbf{1}$ (i.e., the original problem) and minimize (37) over $\{\boldsymbol{\theta}^s\}$ and $\{\rho^s\}$, the least upper bound will obviously be equal to $F(\boldsymbol{\theta})$, attained at $\{\rho^s\}$ given by $\rho^r = 1$ and $\rho^s = 0$ for $s \neq r$. Unfortunately, minimizing over $\{\rho^s\}$ is a non-convex problem.

We did not further follow this interesting research direction.

7. Discrete Markov random fields

As an example of approximative inference in distribution (1), here we assume that (1) is a pairwise MRF.

7.1. MRF as an exponential family

Let the MRF structure be defined by a (finite) set V of variables, where variable $v \in V$ attains states from a finite domain X_v, and a set $E \subseteq \binom{V}{2}$ (thus, (V, E) is an undirected graph).

We show how $(X, I, \boldsymbol{\phi})$ in (1) need to be chosen to represent this MRF. Let $X = \bigtimes_{v\in V} X_v$ be the Cartesian product of the variable domains. A state of variable v is denoted by $x_v \in X_v$. For MRFs, we denote elements of X by boldface letter $\mathbf{x} = (\, x_v \mid v \in V\,)$, rather than by x as for a general exponential family. The set I is given by

$$\begin{aligned} I = \{\, (v,x) \quad &| \; v \in V, \quad x \in X_v \quad \} \\ \cup\, \{\, \{(v,x),(v',x')\} \;&|\; \{v,v'\} \in E,\, x \in X_v,\, x' \in X_{v'}\,\} \end{aligned} \tag{70}$$

In subscripts, the elements of I are denoted by $_{v,x}$ and $_{vv',xx'}$, adopting that $_{vv',xx'}$ is the same as $_{v'v,x'x}$. The functions $\phi_i\colon X \to \{0,1\}$ are indicator functions

$$\phi_{v,y}(\mathbf{x}) = [\![x_v = y]\!] \tag{71a}$$

$$\phi_{vv',yy'}(\mathbf{x}) = \phi_{v,y}(\mathbf{x})\,\phi_{v',y'}(\mathbf{x}) \tag{71b}$$

With this choice of I and $\boldsymbol{\phi}$, we have that

$$\boldsymbol{\theta}\boldsymbol{\phi}(\mathbf{x}) = \sum_{v\in V} \theta_{v,x_v} + \sum_{\{v,v'\}\in E} \theta_{vv',x_v x_{v'}} \tag{72}$$

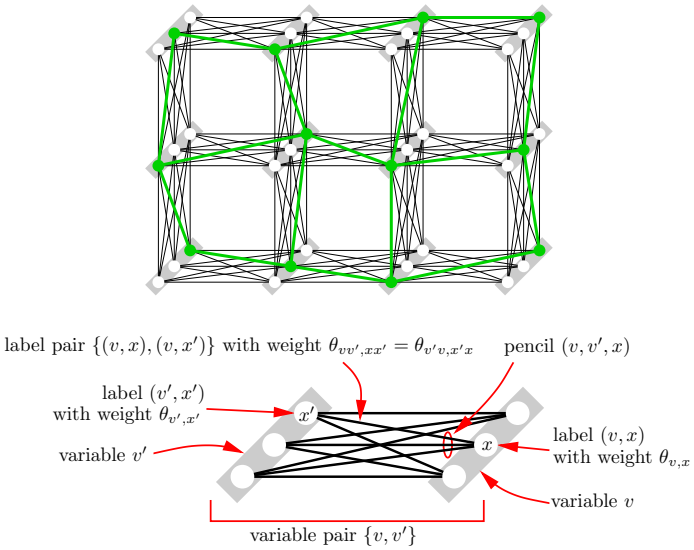


Figure 5. Visualization of pairwise MRF for $|V| = 12$ variables, graph (V,E) being the 3×4 grid graph, and $|X| = 3$ labels. The grey boxes depict the variables (elements of V) and the circles inside them variable states (elements of X_v). Set I is formed by all the circles and edges in the figure. The active elements of I corresponding to an assignment $\mathbf{x}$ is given by $\boldsymbol{\phi}(\mathbf{x})$, thus $\boldsymbol{\phi}(X)$ contains the indicator vectors specifying all valid labelings. An example labeling is depicted in green. The value of $\boldsymbol{\theta}\boldsymbol{\phi}(\mathbf{x})$ equals the sum of weights θ_i sitting on the green nodes and edges.

Distribution (1) becomes a pairwise *Gibbs distribution*, i.e., a distribution defined by a MRF with cliques of size 2 [13]. It can be seen as a special exponential family, characterized by the fact that the mean parameters $m_i(\boldsymbol{\theta})$ coincide with the marginals of $p(\mathbf{x}|\boldsymbol{\theta})$ associated with all variables in V and variable pairs in E. The mean polytope $\operatorname{conv} \boldsymbol{\phi}(X)$ contains all realizable marginal vectors $\boldsymbol{\mu}$ and hence is called the *marginal polytope*. Unlike for general exponential families, the set $\boldsymbol{\phi}(X)$ has certain special properties: each elements of $\boldsymbol{\phi}(X)$ is a vertex of $\operatorname{conv} \boldsymbol{\phi}(X)$ (i.e., $\boldsymbol{\phi}(X)$ is convexly independent) and $\boldsymbol{\phi}(X) = \{0,1\}^I \cap \operatorname{aff} \boldsymbol{\phi}(X)$.

The matrices $\mathbf{A}$ and $\mathbf{B}$, containing the coefficients of affine dependencies (§3.2), are given implicitly in terms of expressions (16) and (17) as follows. The row index sets of $\mathbf{A}$ and $\mathbf{B}$ are $P = \{\, (v,v',x) \mid \{v,v'\} \in E,\ x \in X_v \,\}$ and $Q = V \cup E$, respectively. The expressions $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{B}\boldsymbol{\mu} = \mathbf{1}$ are the marginalization and the normalization constraints

$$\sum_{x'} \mu_{vv',xx'} = \mu_{v,x} \qquad \forall \{v,v'\} \in E,\ x \in X_v \quad (73a)$$

$$\sum_{x} \mu_{v,x} = \sum_{x,x'} \mu_{vv',xx'} = 1 \qquad \forall v \in V,\ \{v,v'\} \in E \quad (73b)$$

Therefore, we have

$$\operatorname{aff} \boldsymbol{\phi}(X) = \left\{ \boldsymbol{\mu} \in \mathbb{R}^I \;\middle|\; \sum_{x'} \mu_{vv',xx'} = \mu_{v,x},\ \sum_{x} \mu_{v,x} = 1 \right\} \quad (74)$$

The reparameterization (17) reads[4]

$$\theta'_{v,x} = \theta_{v,x} - \sum_{v' \in N_v} \alpha_{vv',x} - \beta_v \quad (75a)$$

$$\theta'_{vv',xx'} = \theta_{vv',xx'} \underbrace{+ \alpha_{vv',x} + \alpha_{v'v,x'}}_{-\boldsymbol{\alpha}\mathbf{A}} \underbrace{- \beta_{vv'}}_{-\boldsymbol{\beta}\mathbf{B}} \quad (75b)$$

where $N_v = \{\, v' \mid \{v,v'\} \in E \,\}$ denotes the neighbors of variable v. In subscripts, triplets $(v,v',x) \in P$ are denoted by $_{vv',x}$ and pairs $\{v,v'\} \in E$ by $_{vv'}$. The meaning of homogeneous reparameterization $\boldsymbol{\theta}' = \boldsymbol{\theta} - \boldsymbol{\alpha}\mathbf{A}$ is as follows (see Figure 6 and [28, 10]): for each $(v,v',x) \in P$, subtract number $\alpha_{vv',x}$ from parameter $\theta_{v,x}$ and add the same number to parameters $\{\, \theta_{vv',xx'} \mid x' \in X_{v'} \,\}$. This corresponds to 'a message' in belief propagation literature. Reparameterizations $\boldsymbol{\theta}' = \boldsymbol{\theta} - \boldsymbol{\beta}\mathbf{B}$ correspond to subtracting a constant from each variable v and variable pair $\{v,v'\}$.

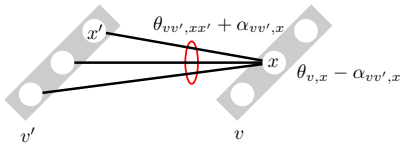


Figure 6. Homogeneous reparameterization.

Next we apply the results from §4 to upper-bound problem (4). In that, we assume that collection Δ is such that all labels of any variable or variable pair either all belong or all do not belong to a subproblem. I.e., we assume there exist $V^s \subseteq V$ and $E^s \subseteq E$ such that

$$\delta^s_{v,x} = [\![v \in V^s]\!] \quad (76a)$$

$$\delta^s_{vv',xx'} = [\![\{v,v'\} \in E^s]\!] \quad (76b)$$

for all $s \in S$. Subproblem s is thus given by a pair (V^s, E^s). To cover the whole problem, we require that $\bigcup_s V^s = V$ and $\bigcup_s E^s = E$. Note that (V^s, E^s) need not be a graph in the strict sense because we do not require that $E^s \subseteq \binom{V^s}{2}$. E.g., we can have $V^s = \emptyset$ and $E^s \neq \emptyset$, in which case the 'graph' (V^s, E^s) has an edge but no nodes.

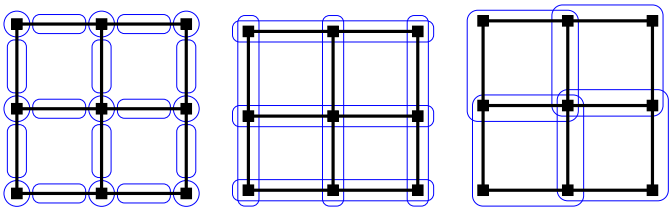
7.2. Arc covering collections

Theorem 5 states on what conditions symbol $\sim$ can be replaced with $=$ in (36). Recall that this would lead to simplified formulations of the optimization problems and optimality conditions and allow to use Algorithm 1 as it is. The assumption of Theorem 5 has a natural interpretation for MRFs, given by Theorem 8. We define an *arc* of graph (V,E) to be an ordered pair $(v,v') \in V \times V$ such that $\{v,v'\} \in E$. We will apply this definition even on a 'graph' (V^s, E^s) with $E^s \not\subseteq \binom{V^s}{2}$.

Theorem 8. *Let for each arc (v,v') of the graph (V,E) there exists a subproblem $s \in S$ such that (v,v') is an arc of (V^s, E^s). Then any $\boldsymbol{\mu} \in \bigcap_{\boldsymbol{\delta} \in \Delta} M_{\boldsymbol{\delta}}$ satisfies $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$.*

Proof. $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$ means that each arc (v,v') of (V,E) satisfies the condition $\sum_{x'} \mu_{vv',xx'} = \mu_{v,x}$. If $\boldsymbol{\mu} \in \bigcap_{\boldsymbol{\delta} \in \Delta} M_{\boldsymbol{\delta}}$, the condition is satisfied for each arc of each subproblem (V^s, E^s)

[4] In [27, 28], numbers $\alpha_{vv',x}$ are denoted by $\phi_{vv',x}$ and have the opposite signs compared to (75).



(a) (b) (c)

Figure 7. Three types of the collection of subproblems Δ.

separately. If each arc of graph (V, E) is covered by at least one subproblem, the condition holds for all arcs of (V, E). ■

The assumption of Theorem 8 means that collection Δ covers not only all nodes and edges of graph (V, E) but *also* all its arcs. In particular, this is true if each (V^s, E^s) is a graph, i.e., if $E^s \subseteq \binom{V^s}{2}$ for $s \in S$. Here are examples of a covered and an uncovered arc (v, v'), respectively:

7.3. Examples of subproblem collections

In this section, we discuss typical examples of the subproblem collection Δ, namely individual nodes and edges, trees, and short cycles. They are depicted in Figure 7.

7.3.1 Individual nodes and edges

Let Δ be the nodes and edges of (V, E), i.e.,

$$\begin{aligned} S &= V \cup E \\ V^v &= \{v\}, \qquad V^{vv'} = \emptyset \\ E^v &= \emptyset, \qquad E^{vv'} = \{v, v'\} \end{aligned}$$

The task (43) simplifies to

$$\min_{\boldsymbol{\theta}' \sim \boldsymbol{\theta}} \Big[\sum_{v \in V} \max_x \theta'_{v,x} + \sum_{\{v,v'\} \in E} \max_{x,x'} \theta'_{vv',xx'} \Big] \tag{77}$$

which can be recognized as minimizing Schlesinger's upper bound [17, 27] over homogeneous reparamerizations of $\boldsymbol{\theta}$. The pair (43)+(44) can be written in a compact way as the following pair of dual linear programs:

$$\begin{aligned} \boldsymbol{\beta}\mathbf{1} &\to \min & \boldsymbol{\theta}\boldsymbol{\mu} &\to \max && \text{(78a)} \\ \boldsymbol{\alpha} &\in \mathbb{R}^P & \mathbf{A}\boldsymbol{\mu} &= \mathbf{0} && \text{(78b)} \\ \boldsymbol{\beta} &\in \mathbb{R}^{V \cup E} & \mathbf{B}\boldsymbol{\mu} &= \mathbf{1} && \text{(78c)} \\ \boldsymbol{\alpha}\mathbf{A} + \boldsymbol{\beta}\mathbf{B} &\geq \boldsymbol{\theta} & \boldsymbol{\mu} &\geq 0 && \text{(78d)} \end{aligned}$$

Clearly, the left-hand program is identical to (77), and the feasible domain of the right-hand program is (80). The polytopes $M_{\boldsymbol{\delta}}$ are just simplices. At optimum, we have $\beta_v = \max_x \theta_{v,x}$ and $\beta_{vv'} = \max_{x,x'} \theta_{vv',xx'}$.

The arcs of (V, E) are not covered by Δ (see Figure 7a) and thus we cannot directly use Algorithm 1. A closely related algorithm that does not require Δ to be arc-covering is *max-sum diffusion* [12, 27]. Its single iteration does the reparameterization of a triplet (v, v', x) such that equality

$$\max_{x'} \theta_{vv',xx'} = \theta_{v,x} \tag{79}$$

becomes satisfied. Repeating this iteration for all triplets converges to a state when (79) is satisfied for all the triplets.

7.3.2 Trees

One possible choice of the collection Δ is a collection of trees, covering the whole graph (V, E). Here, we restate in exponential family terminology two well-known theorems concerning tree collections.

Theorem 9 ([13, 21])**.** *Consider a pairwise MRF such that (V, E) is a tree. Then*

$$\operatorname{conv} \boldsymbol{\phi}(X) = [0, 1]^I \cap \operatorname{aff} \boldsymbol{\phi}(X)$$

Theorem 10 ([21, 20])**.** *For a pairwise arbitrary MRF, let collection Δ contain trees that cover the graph (V, E). Then*

$$M_\Delta = [0, 1]^I \cap \operatorname{aff} \boldsymbol{\phi}(X) \tag{80}$$

Theorem 10 implies that M_Δ does not depend on Δ if Δ contains trees covering the whole graph, as argued in [10].

The polytope $[0, 1]^I \cap \operatorname{aff} \boldsymbol{\phi}(X)$ looks more familiar after substituting for $\operatorname{aff} \boldsymbol{\phi}(X)$ from (74). It is called the *local polytope* in [21, 20, 10].

One choice of Δ is the columns and rows of a grid graph (Figure 7b). Another choice is that each tree is an incident triplet node-edge-node. Both these collections are arc-covering.

7.3.3 4-cycles

Choosing the subproblems Δ as short cycles (as shown in Figure 7c) yields significantly tighter relaxation than with tree subproblems. We have shown this experimentally in [26]. We presented there also a cutting plane strategy for MRFs.

8. Conclusion

We have re-derived the decomposition approach by Wainwright et al. to upper-bounding the modes and partition function of discrete probability distributions of form (1) from general exponential families (i.e., not only from MRF-defined distributions). It has turned out that the task of minimizing the upper bounds, their Lagrange duals, and optimality conditions have simple and natural interpretations (summarized in §4.4), both for non-zero and zero temperature. Moreover, we have given an algorithm to minimize the upper bound, which works for general exponential families (although for zero temperature, the basis functions ϕ_i of the family are restricted to be indicator functions).

What is the advantage of deriving the decomposition approach for general exponential families rather than only for the special case of MRF-defined distributions? From theoretical point of view, this gives a more complete picture on the theory and helps identify which aspects of the decomposition approach are present already on the general level of exponential families and which are specific to MRFs. For instance, it seems so far that arc consistency (= weak tree agreement) has no counterpart in general exponential families. Moreover, the LP pair (78) looks extremely elegant and natural (given only

in terms of the matrices $\mathbf{A}$ and $\mathbf{B}$ defining the affine hull of $\boldsymbol{\phi}(X)$) – but it as well seems to have no counterpart in general exponential families.

From the practical point of view, it is a question to which extent the generalization from MRF-defined distributions to general ones can be useful. If the set X is tractably small, inference can be calculated exactly. Thus, to apply our theory we need X to be intractably large. One example of this is a MRF, where X is a Cartesian product of variable domains. Is there another practically interesting example of distribution (1) with X being combinatorially large but which is not reducible to a MRF? One option is that X is the set of *permutations* – thus we could do statistical inference in *distributions over a set of permutations*. Although we have not detailed this case here, we believe it could be an interesting generalization of the the decomposition approach to MRF inference. On the other hand, it seems quite challenging to find an applications in which distributions over permutations would be useful and which would not at the same time be reducible to an MRF – example problems (their optimization forms) can be various kinds of weighted matching or the weighted Linear Ordering Problem (e.g. [15]). Another choice for a combinatorially large X can be a set of graphs defined by some property, the set of parsing trees of a grammar, etc.

The decomposition approach for zero temperature is closely related to the dual decomposition approach known in optimization. In this approach, the bound for large-scale problems is optimized by subgradient search as in [2, 11], or by a coordinate descent as in [26] and this text. However, if we are interested not only in the modes but also in the partition function and marginals of (1) where X is e.g. a set of permutations, optimization is of no help here and our theory is novel.

A. Proof of Theorems 6 and 7

We will not use the substituion $\mathbf{y} = \boldsymbol{\phi}(x)$ here, i.e., we will use index $x \in X$ rather than $\mathbf{y} \in \mathbf{Y}$.

First, we prove Theorem 6. Before the update (58), we have $F(\boldsymbol{\theta}^s) = \bigoplus_x \boldsymbol{\theta}^s \boldsymbol{\phi}(x)$. After the update, it is easy to verify that

$$F(\boldsymbol{\theta}^s) = \bigoplus_x [\boldsymbol{\theta}^s \boldsymbol{\phi}(x) + \xi^s \phi_i(x)]$$

where ξ^s satisfy (60). We need to show that $\sum_s F(\boldsymbol{\theta}^s)$ after the update is strictly less than before the update.

Denote $\lambda^s(x) = \boldsymbol{\theta}^s \boldsymbol{\phi}(x)$. The rest of the proof is Lemma 1. For clarity, the lemma is formulated in the semiring $(\mathbb{R}_{++}, +, \times)$ rather than in $(\mathbb{R}, \oplus, +)$ and some symbols are abbreviated. Thus, the symbols $\phi_i(x)$, $F(\boldsymbol{\theta}^s)$, $F_i(\boldsymbol{\theta}^s)$, $\lambda^s(x)$, $n_i(\boldsymbol{\theta}^s)$, ξ^s in Theorem 6 correspond respectively to $g(x)$, $\log F^s$, $\log G^s$, $\log \lambda^s(x)$, $\log \mu^s$, $\log \xi^s$ in Lemma 1.

Theorem 7 is proved similarly by translation to Lemma 2.

Lemma 1. *Let $\lambda^s(x) \geq 0$ and $0 \leq g(x) \leq 1$. Denote*

$$F^s = \sum_x \lambda^s(x), \quad G^s = \sum_x \lambda^s(x) g(x), \quad \mu^s = \frac{G^s}{F^s} \tag{81}$$

Let $\xi^s \geq 0$ be such that

$$\sum_s \xi^s \mu^s < \sum_s \mu^s \tag{82}$$

Then

$$\prod_s \sum_x \lambda^s(x) (\xi^s)^{g(x)} < \prod_s F^s \tag{83}$$

Proof. The proof proceeds by bounding the left-hand side of (83) twice, which translates the inequality (83) into (82).

Any $\xi > 0$ and any $0 \leq g \leq 1$ satisfy[5] $\xi^g \leq 1 + (\xi - 1)g$ because the right-hand side is the tangent line of the concave function $\xi \mapsto \xi^g$ at $\xi = 1$. Applying this inequality on the factors on the left-hand side of (83) yields

$$\begin{aligned} \sum_x \lambda^s(x) (\xi^s)^{g(x)} &\leq \sum_x \lambda^s(x) [1 + (\xi^s - 1) g(x)] \\ &= F^s + (\xi^s - 1) G^s \end{aligned}$$

Therefore, it suffices to prove that

$$\prod_s [F^s + (\xi^s - 1) G^s] < \prod_s F^s$$

which after dividing each factor by F^s reads

$$\prod_s [1 + (\xi^s - 1)\mu^s] < 1 \tag{84}$$

By the arithmetic-geometric average inequality, the inequality (84) is implied by

$$\sum_s [1 + (\xi^s - 1)\mu^s] < |S|$$

which is easily simplified to (82). ∎

Lemma 2. *Let $g(x) \in \{0, 1\}$ and let $\sum$ be replaced with $\max$ in (81), (82) and (83). Then Lemma 2 remains true, with the exception that inequality (83) may not be strict.*

Proof. Suppose that $\sum$ has been replaced with $\max$. Denoting $\alpha = \{\, x \mid g(x) = 1 \,\}$, we can write $G^s = \max_{x \in \alpha} \lambda^s(x)$ and

$$\max_x \lambda^s(x) (\xi^s)^{g(x)} = \max \left\{ \xi^s G^s, \max_{x \notin \alpha} \lambda^s(x) \right\} \tag{85}$$

Clearly, $\max_{x \notin \alpha} \lambda^s(x) \leq F^s$. Moreover, $\xi^s G^s \leq F^s$, which follows from (82) and from the obvious fact that $\mu^s \leq 1$. Thus, the expression (85), and hence each factor on the left-hand side of (83), is not greater than F^s. ∎

References

[1] Shun-Ichi. Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford University Press, 1993.

[2] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1999.

[3] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] L. D. Brown. *Fundamentals of statistical exponential families: With applications in statistical decision theory*. Institute of Mathematical Statistics, Hayworth, CA, USA, 1986.

[5] I. Csiszár. A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling. *Annals Statist.*, 17:1409–1413, 1989.

[6] J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43:1470–1480, 1972.

[5] The subscript ξ^g denotes power, unlike in ξ^s where it denotes an index.

[7] S. E. Fienberg. An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, 41(3):907–917, 1970.

[8] Stéphane Gaubert. Methods and applications of (max,+) linear algebra. Technical Report 3088, Institut national de recherche en informatique et en automatique (INRIA), 1997.

[9] Jason K. Johnson, Dmitry M. Malioutov, and Alan S. Willsky. Lagrangian relaxation for MAP estimation in graphical models. In *Allerton Conf. Communication, Control and Computing*, 2007.

[10] Vladimir Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(10):1568–1583, 2006.

[11] N. Komodakis and N. Paragios. Beyond loose LP-relaxations: Optimizing MRFs by repairing cycles. In *European Conf. on Computer Vision (ECCV)*, 2008.

[12] V. A. Kovalevsky and V. K. Koval. A diffusion algorithm for decreasing energy of max-sum labeling problem. Glushkov Institute of Cybernetics, Kiev, USSR. Unpublished, approx. 1975.

[13] S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[14] Grigori L. Litvinov. The Maslov dequantization, idempotent and tropical mathematics: A brief introduction. *Mathematical Sciences*, 2006. Available from http://arXiv.org.

[15] Alantha Newman and Santosh Vempala. Fences are futile: On relaxations for the linear ordering problem. In *8th Intl. IPCO Conf. on Integer Programming and Combinatorial Optimization*, pages 333–347, London, UK, 2001. Springer-Verlag.

[16] Pradeep Ravikumar, Alekh Agarwal, and Martin J. Wainwright. Message-passing for graph-structured linear programs: proximal projections, convergence and rounding schemes. In *Intl. Conf. on Machine Learning (ICML)*, pages 800–807. ACM, 2008.

[17] M. I. Shlezinger. Syntactic analysis of two-dimensional visual signals in noisy conditions. *Cybernetics and Systems Analysis*, 12(4):612–628, 1976. Translation from Russian: Sintaksicheskiy analiz dvumernykh zritelnikh signalov v usloviyakh pomekh, Kibernetika, vol. 12, no. 4, pp. 113-130, 1976.

[18] M. Wainwright, T. Jaakkola, and A. Willsky. MAP estimation via agreement on (hyper)trees: message passing and linear programming approaches. In *Allerton Conf. on Communication, Control and Computing*, 2002.

[19] M. Wainwright, T. Jaakkola, and A. Willsky. Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Trans. Information Theory*, 49(5):1120–1146, 2003.

[20] M. Wainwright, T. Jaakkola, and A. Willsky. MAP estimation via agreement on (hyper)trees: message passing and linear programming approaches. *IEEE Trans. Information Theory*, 51(11):3697–3717, 2005.

[21] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, UC Berkeley, Dept. of Statistics, 2003.

[22] Martin J. Wainwright, Tommi Jaakkola, and Alan S. Willsky. A new class of upper bounds on the log partition function. In *18th Conference in Uncertainty in Artificial Intelligence (UAI), University of Alberta, Edmonton, Alberta, Canada*, pages 536–543. Morgan Kaufmann, August 2002.

[23] Martin J. Wainwright, Tommi Jaakkola, and Alan S. Willsky. A new class of upper bounds on the log partition function. *IEEE Trans. Information Theory*, 51(7):2313–2335, 2005.

[24] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[25] Yair Weiss, Chen Yanover, and Talya Meltzer. MAP estimation, linear programming and belief propagation with convex free energies. In *Conf. Uncertainty in Artificial Intelligence (UAI)*, 2007.

[26] T. Werner. High-arity interactions, polyhedral relaxations, and cutting plane algorithm for soft constraint optimisation (MAP-MRF). In *Computer Vision and Pattern Recognition (CVPR) Conf., Anchorage, USA*, June 2008.

[27] Tomáš Werner. A linear programming approach to max-sum problem: A review. Technical Report CTU–CMP–2005–25, Center for Machine Perception, Czech Technical University, December 2005.

[28] Tomáš Werner. A linear programming approach to max-sum problem: A review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(7):1165–1179, July 2007.

[29] Tomas Werner. Marginal consistency: Unifying constraint propagation on commutative semirings. In *Intl. Workshop on Preferences and Soft Constraints (co-located with Conf. on Principles and Practice of Constraint Programming)*, pages 43–57, September 2008.

[30] Tomáš Werner and Alexander Shekhovtsov. Unified framework for semiring-based arc consistency and relaxation labeling. In *12th Computer Vision Winter Workshop, St. Lambrecht, Austria*, pages 27–34. Graz University of Technology, February 2007.